KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

# Unsupervised Algorithms for Cross-Lingual Text Analysis, Translation Mining, and Information Retrieval

**Ivan Vulić**

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

June 2014

# Unsupervised Algorithms for Cross-Lingual Text Analysis, Translation Mining, and Information Retrieval

**Ivan VULIĆ**

Examination Committee:
Prof. dr. ir. Paul Van Houtte, chair
Prof. dr. Marie-Francine Moens, supervisor
Prof. dr. ir. Hendrik Blockeel
Prof. dr. Danny De Schreye
Prof. dr. ir. Hugo Van hamme
Dr. Kris Heylen

Prof. dr. Stephen Clark
(University of Cambridge)
Prof. dr. ir. Hermann Ney
(RWTH Aachen University)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

June 2014

D/2014/7515/64
ISBN 978-94-6018-840-4

# Acknowledgments

I could use so many different metaphors to describe my Ph.D. experience. A “roller coaster” may work just fine. Or how about “an unexpected journey”? Or maybe “2009-2014: A Belgian Odyssey”? Or, maybe the name of that ancient TV show which suits perfectly the occasion - "Boy Meets World"? To wrap it up, the choice of doing a Ph.D. in Leuven was one of the most important threads in my personal “random weaving of destiny” (I forewarn you that I sometimes tend to use big words when I get to writing) called life. And the weaving of this thread included many different people who provided material, ideas, designs, needles, etc. Thanking all of you in only a few lines won't make justice to all of you. If I happen to forget someone, I apologize in advance!

First of all, I would like to thank Sien Moens, my Ph.D. promoter and advisor, for, well, for many things. First, for providing me an opportunity to pursue a Ph.D. and having trust in me, although the whole trust in the beginning relied on my promise that “I'm able to learn fast enough”. I could really say that I started this Ph.D. as a *tabula rasa* - I had no idea how to do proper scientific research or write scientific papers, I never had a course in natural language processing or information retrieval, I was supposed to build algorithms that translate from French to Dutch which were not, to put it that way, the strongest points in my resume at the time, and, even more, my trip to Belgium was the first flight in my life. Thanks to you, Sien, I have learned all this - how to do proper and thorough scientific research, how to be systematic and how to write papers, and I even managed to attend conferences all around the world (traveling by plane!), from Seattle or New York all the way to Mumbai, Dubai

or Australia (the “Boy Meets World” analogy is quite straightforward here). I would also like to thank you for letting me pursue my own ideas independently, but providing guidance whenever I strayed away or was stubborn too much to accept that I had been wrong. Dankjewel, Sien!

I would also like to thank Kris Heylen, who was the coordinator of the TermWise project. I learned a lot from our work together and our discussions. I apologize for being too much a computer scientist and too little a linguist sometimes. Thank you also for agreeing to serve in my examination committee. I would also like to express my gratitude to professors Hermann Ney and Stephen Clark for serving as the external members of my examination committee, as well as the other members of the committee, professors Hendrik Blockeel, Danny De Schreye and Hugo Van hamme, for taking their time to evaluate this thesis, provide valuable feedback and ideas for future research. Also thanks to professor Paul Van Houtte for serving as the chairman of the committee.

One big thank you goes to all my colleagues and co-authors at LIIR, past and present. It was a fun and pleasant experience to be here all these years and to share all ups and downs of doing a Ph.D. with you. Thanks goes to Oleksandr, who was my desk mate for more than four years, and often the only living being between me and the outer world (I mean, literally, he sits by the window!). Thanks, Wim, for your guidance during the early stage of my Ph.D., and for introducing me to the world of topic modeling. Thanks, Steve, for always finding time to listen to my ideas, to read drafts of my papers, and to provide advice on how to improve them. I would also like to thank Jan for translating the abstract to Dutch.

One big thank you goes to the research lab at my home university in Zagreb (especially Goran), and professors Bojana Dalbelo-Bašić and Jan Šnajder for directing me towards doing a Ph.D. when I bursted into their office almost 5 years ago asking for advice, and also for hosting me in their lab for almost two months while I was full-time writing this manuscript last year. Hvala puno!

One huge bag full of “thank yous” goes to all my friends, back home, in Zagreb, in Leuven, in Belgrade, all over the world. To all my “old country” (and now almost ex-Leuven) crew, in no particular order: Bogi, Mire, Ivan, Tićma, Duško & Milica, Aćko, Vukov & Sneža, Jopa The Face, Miki & Dara, Ilija, Dušan, Ljilja, Meto, Damjan, thank you for all the basketball games, all the barbecues, all the nights out, all the travels, all the beers and all the coffees; thank you for being there to just, you know, hang out and have good time! Hvala vam svima koliko vas ima! ili “Moji su drugovi biseri rasuti...”. Thank you, Sofija, for showing me every single bar in Belgrade, for being honest, fun and being yourself. One big thanks goes to Boran, for all the pancakes and “crash bash”, for laughing out loud while watching “Ljubav je na selu”, for surviving snowy

winters and rainy summers with video games, pizzas, bad jokes, listening to "Kuzma and Shaka Zulu", and waiting for "điran" to blossom. To Zaher (or you can just call him Steve) & Ivana, thanks for all the "rectangular conversations" (i.e., Google chats), and for your long distance friendship, and for always letting me crash at your place (even when this "crashing at your place" sometimes meant living with me around for more than two months). Thank you, Anđelo, for being a great flatmate and friend all these years. I hope I have not caused any permanent damage to your ears and brain with my singing and guitar playing. And, also, one really special thank you goes to Marina & Bane, described once (and repeated here) as "najveći carevi Leuvena". Whenever I felt blue or homesick or needed a drink and some company, or needed someone to talk to after I had come back to this rainy Belgium in the summer and felt utterly depressed, they were here, in Leuven, with a cup of coffee, a plate of beans or chicken wings, a kind word, a new joke, a new reason to feel a bit better. Hvala na svemu!

I would also like to thank my uncles, aunts, cousins, and especially my grandmas and grandpas (to baba i dido, I wish you a lot more "sriće i zdravlja"!), for being around since I was a little kid until this very moment in time. I have to apologize, but you are in too many to write a personal note to every one of you.

I owe the most to my family. All their support and love cannot fit into these few lines of gratitude. They are the ones who get to see (usually via Skype) and deal with "the yin and yang of me", my best, but also my worst days. My mom and dad were always there for me, from the early days, directing me towards important things in life, teaching me valuable lessons and providing encouragement and unconditional support. Hvala, mama! Hvala, tata! To Josipa and Katarina, I do not have any special words for you with all these years of growing up with or (more often) without your big brother, you know everything already (even if it does not always look that way! :)). And, who knows, maybe Luka, my first nephew, who is about to enter this world a few days after this text, will get to read these lines in a few years time?

And finally, I would like to thank Anita, my dearest Pake Pakasto. For all the universes and all the planets, for all the worlds and all the time lines, for all the jungles and all the deserts, for all the savannahs and all the baobab trees, for all the mountains and all the valleys, for all the oceans and all the rivers, for all the dusks and all the dawns, for all the words and all the thoughts we have been walking over, through and across, alone and together, some hidden in the dark, some bright and shiny, some obscure, some frightening, some brilliant, some inspiring, some astonishing, some exceptional... And just wonderful.

Ivan Vulić
Leuven, June 2014

# Abstract

With the ongoing growth of the global network and information influx in today's increasingly connected world, more and more content becomes readily available in a plethora of different languages, dialects, unofficial and community languages. Considering the large amount of multilingual data which are typically unstructured but thematically aligned and comparable, there is a pressing need to build unsupervised algorithms which can deal with such multilingual data, and address the problems of meaning, translation and information retrieval in multilingual settings.

The thesis has four major contributions in the research fields of data mining, natural language processing and information retrieval. First, we present and describe a full overview of the newly developed multilingual probabilistic topic modeling (MuPTM) framework for mining multilingual data. The framework is utilized to induce high-level language-independent representations of textual information (e.g., words, phrases and documents). Second, we propose a new statistical framework for inducing bilingual lexicons (i.e., addressing the problem of translation) from parallel data that is based on the novel paradigm of sub-corpora sampling. Third, we introduce a new statistical framework for modeling cross-lingual semantic similarity (i.e., addressing the problem of meaning) and inducing bilingual lexicons (i.e., the problem of translation) from comparable data. Here, we make a series of contributions to the field of (multilingual) natural language processing and its sub-field of distributional semantics by (i) proposing a series of MuPTM-based models of cross-lingual semantic similarity, (ii) designing an algorithm for detecting only highly reliable

translation pairs from noisy multilingual environments, (iii) proposing a new language pair independent cross-lingual semantic space that relies on the concept of semantic word responding, (iv) presenting a new bootstrapping approach to cross-lingual semantic similarity and bilingual lexicon extraction, and (v) proposing a new context-sensitive framework for modeling semantic similarity. Fourth, we propose a new probabilistic framework for cross-lingual and monolingual information retrieval (i.e., tackling the problem of information retrieval) which relies on MuPTM-based text representations.

All proposed models are unsupervised and language pair independent in their design. Consequently, that makes them potentially applicable to many language pairs. The proposed models have been evaluated with a variety of language pairs, and we show that they advance state-of-the-art in their respective fields. Due to their unsupervised and language pair independent nature, the presented models exhibit a solid potential for future research and other applications that deal with different official and unofficial languages, dialects and different idioms of the same language.

# Beknopte samenvatting

Met de toenemende instroom van informatie in onze steeds verbondener wordende wereld komt er meer en meer inhoud beschikbaar, in een verscheidenheid aan talen, dialecten, onofficiële en community talen. Wanneer we kijken naar de grote hoeveelheid aan informatie die beschikbaar is in meerdere talen, maar zelden een exacte vertaling is, zien we dat er een grote nood is aan ongesuperviseerde algoritmen, die kunnen omgaan met deze ongestructureerde meertalige gegevens, en zich richten op het achterhalen van de betekenis, het vertalen, en het zoeken in een meertalige omgeving.

Deze thesis levert vier grote bijdragen voor de onderzoeksdomeinen data mining, natuurlijke taalverwerking, en information retrieval. Eerst geven we een gedetailleerd overzicht van het nieuw ontwikkelde meertalige probabilistisch *topic modeling framework* (MuPTM) voor het mijnen van meertalige data. Dit framework gebruiken we om een taal-onafhankelijke voorstelling van tekstuele informatie (bv. woorden, woordgroepen, en documenten) op een hoog niveau te induceren. Als tweede bijdrage stellen we een nieuw statistisch framework voor, dat bilinguale lexicons opstelt (met als toepassing het vertalen van tekst) uit parallelle data die gebaseerd is op het nieuwe paradigma van sub-corpora sampling. Als derde introduceren we een nieuwe statistisch framework voor het modelleren van cross-linguale semantische similariteit (i.e., om de betekenis van tekst te achterhalen) en om bilinguale lexicons te induceren uit vergelijkbare data. Hierbij maken we een reeks van bijdragen tot het domein van (meertalige) natuurlijke taalverwerking en het subdomein van distributionele semantiek met (i) modellen gebaseerd op MuPTM voor cross-linguale semantische

similariteit, (ii) een nieuw algoritme dat enkel zeer betrouwbare vertalingen vindt in meertalige omgevingen met veel ruis, (iii) een nieuwe cross-linguale semantische ruimte, onafhankelijk van het taalpaar, die bouwt op semantische woord antwoorden, een concept uit de cognitieve wetenschap, (iv) een nieuwe bootstrapping methode voor cross-linguale semantische similariteit en bilinguale lexicon extractie, en (v) een nieuw context-gevoelig framework voor het modelleren van semantische similariteit. Als vierde bijdrage presenteren we een nieuw probabilistisch framework voor cross-linguale en monolinguale information retrieval, dat terugvalt op MuPTM gebaseerde tekstvoorstellingen.

Alle voorgestelde modellen zijn ongesuperviseerd, en in hun ontwerp onafhankelijk van de taalparen. Bijgevolg zijn ze toepasbaar op veel combinaties van twee talen. De voorgestelde modellen zijn geëvalueerd met een grote verscheidenheid aan taalparen, en we tonen aan dat ze de huidige beste methoden verbeteren in hun respectievelijke domeinen. Omdat ze ongesuperviseerd en onafhankelijk zijn van de taalparen, tonen de modellen een duidelijk potentieel voor verder onderzoek en andere toepassingen waarbij er moet omgegaan worden met meerdere officiële en onofficiële talen, dialecten, en verschillende idiomen van dezelfde taal.

# Kratki sažetak

Posljedično s postojanim neposustajućim rastom globalne mreže i sve većim priljevom informacija u današnjem sve povezanijem svijetu, sve je više sadržaja na raspolaganju u mnoštvu različitih službenih i neslužbenih jezika, narječja i idiomatskih dijalekata specifičnih za pojedine grupe ili udruženja. Uzevši u obzir te ogromne količine višejezičnih podataka koji su najčešće nestrukturirani, ali unatoč tome često tematski poravnati i usporedivi, stvara se sve jača potreba za ostvarivanjem nenadziranih algoritama koji su u stanju procesirati i analizirati višejezične podatke te ponuditi rješenja za temeljne probleme značenja, prevođenja te pretrage i dohvata informacija u višejezičnim okruženjima.

Ova disertacija ostvaruje četiri važna doprinosa istraživačkim područjima dubinske analize podataka, obrade prirodnog jezika i dohvata informacija. Kao prvi doprinos, predstavljena je skupina nedavno nastalih višejezičnih vjerojatnosnih tematskih modela koja se koristi za analizu višejezičnih podataka. Ovi modeli koriste se za konstrukciju visokorazinskih jezično neovisnih prikaza tekstualnih podataka (npr., riječi, fraza i dokumenata). Kao drugi doprinos, predstavljen je nov statistički radni okvir za automatsku ekstrakciju dvojezičnih rječnika iz usporednih korpusa. Ovaj doprinos vezan je uz problem prevođenja. Kao treći doprinos, uveden je i predstavljen nov statistički radni okvir za modeliranje međujezične semantičke sličnosti (problem značenja) i automatsku ekstrakciju dvojezičnih rječnika iz usporedivih korpusa (problem prevođenja). U sklopu ovog doprinosa, ostvaren je i niz važnih doprinosa istraživačkom području obrade prirodnog jezika i njegovom potpodručju distribucijske semantike: (i)

predstavljen je niz modela međujezične semantičke sličnosti koji se temelje na višejezičnim vjerojatnosnim tematskim modelima, (ii) ostvaren je algoritam za otkrivanje visokopouzdanih prijevodnih parova iz nepouzdanih usporedivih podataka, (iii) predložen je nov međujezični semantički prostor neovisan o izabranom paru jezika koji se temelji na ideji semantičkog uzvraćanja, (iv) predstavljen je nov iterativni samopodržavajući pristup modeliranju semantičke sličnosti i ekstrakciji dvojezičnih rječnika, i (v) predložen je nov radni okvir za modeliranje kontekstno ovisne semantičke sličnosti. Konačno, kao čevrti doprinos, predložen je nov vjerojatnosni radni okvir za ostvarivanje modela dohvata informacija (problem pretrage i dohvata informacija) koji se temelji na prikazima riječi i dokumenata temeljenim na višejezičnim vjerojatnosnim tematskim modelima.

Svi predloženi modeli su nenadzirani i u svojem ostvaraju neovisni o izabranom paru jezika. Kao posljedica, omogućena je primjena predloženih modela na mnoštvo parova jezika. Predloženi modeli evaluirani su na raznovrsnim parovima jezika i pokazano je da modeli postižu rezultate iznad performansi trenutno najboljih modela u pojedinim istraživačkim područjima. Zbog svojih svojstava nenadgledanosti i jezične neovisnosti predstavljeni modeli pokazuju velik potencijal u budućem istraživanju i novim primjenama u kojima će biti korišteni različiti službeni i neslužbeni jezici, idiomi i narječja.

# List of Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| BLE | Bilingual Lexicon Extraction |
| BiLDA | Bilingual Latent Dirichlet Allocation |
| CAT | Computer-Assisted Translation |
| CL | Computational Linguistics |
| CLEF | Conference and Labs of the Evaluation Forum |
| CLIR | Cross-Lingual Information Retrieval |
| CRM | Cross-Lingual Relevance Model |
| DM | Document Model |
| EM | Expectation-Maximization |
| IR | Information Retrieval |
| LDA | Latent Dirichlet Allocation |
| LLR | Log-Likelihood Ratio |
| LM | Language Model |
| LSA | Latent Semantic Analysis |
| MAP | Mean Average Precision |
| MAPE | Maximum A Posteriori Estimation |
| MIR | Multilingual Information Retrieval |
| MLE | Maximum Likelihood Estimation |
| MRM | Monolingual Relevance Model |
| MRR | Mean Reciprocal Rank |
| MoPTM | Monolingual Probabilistic Topic Modeling |
| MuPTM | Multilingual Probabilistic Topic Modeling |
| NLP | Natural Language Processing |
| PLSA | Probabilistic Latent Semantic Analysis |
| PMI | Pointwise Mutual Information |
| POS | Part-Of-Speech |
| PPMI | Positive Pointwise Mutual Information |
| PTM | Probabilistic Topic Modeling |
| PolyLDA | Polylingual Latent Dirichlet Allocation |
| RM | Relevance Model |
| SMT | Statistical Machine Translation |

# List of Symbols

| Introduced in Fundamentals | |
|---|---|
| $\sim$ | Distributed according to |
| $B(a, b)$ | Beta function |
| $\Gamma(a)$ | Gamma function |
| $cdf$ | Cumulative distribution function |
| $f$ | Probability function |
| $P$ | Probability measure |
| $\mathbf{p}$ | Vector of success parameters |
| $p_s, p_i$ | Success parameters |
| $pdf$ | Probability density function |
| $pmf$ | Probability mass function |
| $\mathcal{W}$ | Probability space |
| $w, w_i$ | Word in a text document |
| $\mathcal{X}$ | Event space |
| $X$ | Stochastic variable |
| $x$ | Outcome of a random variable |

| Introduced in Part I | |
|---|---|
| $\alpha, \alpha_i$ | Dirichlet prior on distribution $\theta$ |
| $\beta$ | Dirichlet prior on distributions $\phi$ and $\psi$ |
| $\phi_i$ | Per-topic word distribution in language $L_i$ |
| $\phi$ | Per-topic word distribution in the source language $L_S$ |
| $\psi$ | Per-topic word distribution in the source language $L_T$ |
| $\theta$ | Per-document topic distribution |
| $\mathcal{C}$ | Multilingual text corpus |
| $\mathcal{C}_i$ | Collection of documents from $\mathcal{C}$ given in language $L_i$ |
| $d^S, d_i^S$ | Source language document |
| $d^T, d_i^T$ | Target language document |
| $K$ | Number of latent cross-lingual topics/concepts in $\mathcal{Z}$ |
| $\mathcal{L}$ | Set of $l$ languages |
| $L_i$ | $i$-th language in the set $\mathcal{L}$ |
| $L_S$ | Source language |
| $L_T$ | Target language |

| | |
|---|---|
| $n^S_{j,k}$ | Count of the number of word tokens in document $d^S_j$ assigned to topic $z_k$ |
| $n^S_{j,k,\neg i}$ | Count of the number of word tokens in document $d^S_j$ assigned to topic $z_k$ excluding word $w^S_{ji}$ at position $i$ |
| $n^T_{j,k}$ | Count of the number of word tokens in document $d^T_j$ assigned to topic $z_k$ |
| $n^T_{j,k,\neg i}$ | Count of the number of word tokens in document $d^T_j$ assigned to topic $z_k$ excluding word $w^T_{ji}$ at position $i$ |
| $V^i$ | Vocabulary of language $L_i$ |
| $V^S$ | Source language vocabulary |
| $V^T$ | Target language vocabulary |
| $\|V^S\|$ | Size of the source language vocabulary |
| $\|V^T\|$ | Size of the target language vocabulary |
| $v^S_{k,w_{ji},\neg}$ | Count of the number of times a word which occurs at position $i$ ($w^S_{ji}$) gets assigned a topic $z_k$ in the corpus $\mathcal{C}_S$ not counting $w^S_{ji}$ |
| $v^T_{k,w_{ji},\neg}$ | Count of the number of times a word which occurs at position $i$ ($w^T_{ji}$) gets assigned a topic $z_k$ in the corpus $\mathcal{C}_T$ not counting $w^T_{ji}$ |
| $\mathbf{w}^S$ | All word tokens in the source language corpus $\mathcal{C}_S$ |
| $\mathbf{w}^T$ | All word tokens in the source language corpus $\mathcal{C}_T$ |
| $\mathbf{w}^S_{\neg ji}$ | All word tokens in $\mathcal{C}_S$ excluding $w^S_{ji}$ |
| $\mathbf{w}^T_{\neg ji}$ | All word tokens in $\mathcal{C}_T$ excluding $w^T_{ji}$ |
| $w^S_{ji}$ | Word at position $i$ in document $d^S_j$ |
| $w^T_{ji}$ | Word at position $i$ in document $d^T_j$ |
| $w^S, w^S_i$ | Source language word from vocabulary $V^S$ |
| $w^T, w^T_i$ | Target language word from vocabulary $V^T$ |
| $\mathbb{Y}$ | Bayesian network |
| $Y$ | Stochastic variable from $\mathbb{Y}$ |
| $\mathcal{Z}$ | Set of latent cross-lingual topics/concepts |
| $\mathbf{z}^S_j$ | All topic assignments for document $d^S_j$ |
| $\mathbf{z}^T_j$ | All topic assignments for document $d^T_j$ |
| $\mathbf{z}^S_{\neg ji}$ | All topic assignments for document $d^S_j$ excluding for $w^S_{ji}$ |
| $\mathbf{z}^T_{\neg ji}$ | All topic assignments for document $d^T_j$ excluding for $w^T_{ji}$ |
| $z_k$ | $k$-th latent cross-lingual topic/concept from set $\mathcal{Z}$ |
| $z^S_{ji}$ | topic assignment for a word at position $i$ in document $d^S_j$ |
| $z^T_{ji}$ | topic assignment for a word at position $i$ in document $d^T_j$ |

Introduced in Part II

| | |
|---|---|
| $\Omega$ | Number of aligned items (e.g., sentences) in the corpus $\mathcal{C}$ |
| $\omega$ | Number of aligned items (e.g., sentences) in the sub-corpus $\mathcal{SC}$ |
| $C(w^S_1, w^T_2)$ | Co-occurrence count of two words $w^S_1$ and $w^T_2$ |
| $C(w^S_1)$ | Frequency count of a word $w^S_1$ |
| $\mathcal{F}$ | Set of features used to extract potential translation pairs |
| $F_f$ | A feature from $\mathcal{F}$: Minimum frequency threshold |
| $F_i$ | A feature from $\mathcal{F}$: Minimum number of aligned item pairs |
| $GTC(w^S_1)$ | Correct translation of $w^S_1$ as given by ground truth $G$ |
| $I^S_i$ | Item $i$ (e.g., sentence) on the source side of the corpus $\mathcal{C}$ or $\mathcal{SC}$ |
| $I^T_i$ | Item $i$ (e.g., sentence) on the target side of the corpus $\mathcal{C}$ or $\mathcal{SC}$ |
| $os_{t_{ij}}$ | Score assigned to a potential translation pair $t_{ij}$ |
| $rank(GTC(w^S_1))$ | Rank of the correct translation of $w^S_1$ in the list of its candidate translations |
| $\mathcal{SC}$ | Sub-corpus of the corpus $\mathcal{C}$ |
| $t_{ij}$ | Potential translation pair |
| $weight_\omega$ | Weight assigned to a translational equivalence detected in the sub-corpus $\mathcal{SC}$ of size $\omega$ |

| Introduced in Part III | |
|---|---|
| $\Delta, \Delta_0, \Delta_{min}$ | Threshold: values and parameters in the SelRel algorithm |
| $\gamma, \lambda, \lambda_1, \lambda_2$ | Interpolation parameters |
| $Acc_M$ | Top $M$ accuracy |
| $B$ | Number of new dimensions to be added in one iteration of the bootstrapping procedure |
| $CF(w_1^S, TC(w_1^S))$ | Confidence score of a translation pair $(w_1^S, TC(w_1^S))$ |
| $Con(w_i^S)$ | Context of a word $w_i^S$ |
| $c_i$ | Context feature |
| $cw_j^S$ | Context word from the context $Con(w_i^S)$ |
| $dec_\Delta$ | Threshold decrementing step |
| $K'$ | Number of features/topics after topic space pruning |
| $L_o, L_f$ | One-to-one bilingual lexicons |
| $M_0, M_c, M_{cmax}$ | Search space depth: values and parameters in the SelRel algorithm |
| $N$ | Number of features/dimensions in the context vector |
| $P'(z_k\|w_i^S)$ | Modulated conditional topic probability for topic $z_k$ and word $w_i^S$ |
| $\text{pmi}_{ik}^S$ | PMI score for word $w_i^S$ associated with context feature $c_k$ |
| $RL(w_1^S)$ | Ranked list for $w_1^S$ |
| $RL_M(w_1^S)$ | Ranked list for $w_1^S$ pruned at rank $M$ |
| $RRL(w_1^S)$ | "Re-ranked" ranked list for $w_1^S$ |
| $SF$ | Similarity function |
| $sc_i^S(c_n)$ | Score associated with the $n$-th feature/dimension of the vector $vec(w_1^S)$ |
| $sim(w_1^S, w_2^T)$ | Semantic similarity score between two words |
| $\text{smoothedpmi}_{ik}^S$ | Smoothed PMI score for word $w_i^S$ associated with context feature $c_k$ |
| $\mathcal{T}, \mathcal{T}_o, \mathcal{T}_s$ | Dimensions of the bootstrapped cross-lingual vector space |
| $\mathcal{TC}(w_1^S)$ | Sense inventory/Inventory of translation candidates of $w_1^S$ |
| $TC(w_1^S)$ | Translation candidate for $w_1^S$ |
| TF-IDF | Term frequency - inverse document frequency |
| TTF-ITF | Term-topic frequency - inverse topic frequency |
| $vec(w_1^S)$ | Context (or feature) vector for $w_1^S$ |
| $w_{1,i}^T$ | Target language word at position $i$ in $RL(w_1^S)$ |
| $w_{i,high}^T$ | Best scoring word from $RRL(w_i^S)$ |

| Introduced in Part IV | |
|---|---|
| $\delta_1, \delta_2, \delta_3, \delta_4$ | Interpolation parameters |
| $\mu$ | Parameter of the Dirichlet prior in the Dirichlet smoothing scheme |
| $A_j = (A_j^S, A_j^T)$ | Aligned Wikipedia article pair given in languages $L_S$ and $L_T$ |
| $Coll^S$ | Collection of Wikipedia articles given in language $L_S$ |
| $cf_{\mathcal{DC}^T}(q_i^S)$ | Corpus frequency: Number of occurrences of $q_i^S$ in the entire document collection $\mathcal{DC}^T$ |
| $\mathcal{DC}^T$ | Target document collection in language $L_T$ |
| $len$ | Query length |
| $DRank(Q^S)$ | Ranking of documents from the target document collection according to their relevance to $Q^S$ |
| $N_{d_j^T}$ | Length of document $d_j^T$ given in the number of word tokens |
| $\mathcal{Q}^S$ | Query collection in language $L_S$ |
| $Q^S$ | Query issued in language $L_S$ |
| $q_i^S$ | Query term from $Q^S$ |
| $R_Q$ | Set of documents relevant to a user's query $Q$ |
| $Ref^S$ | Monolingual reference corpus given in language $L_S$ |
| $tf_{d_j^T}(q_i^S)$ | Term frequency: Number of occurrences of $q_i^S$ in document $d_j^T$ |

# Contents

**Abstract** v

**List of Abbreviations** xi

**List of Symbols** xiii

**Contents** xvii

**List of Figures** xxvii

**List of Tables** xxxi

**List of Algorithms** xxxv

**1 Introduction** 1

1.1 Motivation and Goals . . . . . 4

1.2 Contributions . . . . . 6

1.3 Structure of the Thesis . . . . . 7

**2 Fundamentals** **11**

2.1 Introduction . . . . . 11

2.2 A (Too) Short Introduction to Probability Theory and Bayesian Modeling . . . . . 12

2.2.1 Basic Concepts of Probability Theory . . . . . 12

2.2.2 Important Probability Distributions . . . . . 14

2.2.3 Sampling from a Distribution . . . . . 17

2.2.4 Bayesian Modeling and Generative Models . . . . . 18

2.2.5 Latent Variables and their Estimation . . . . . 19

2.3 Statistical Language Modeling . . . . . 21

2.4 Text Preprocessing . . . . . 22

2.5 Conclusion . . . . . 24

**I Multilingual Text Mining** **25**

**3 Multilinguality and Multilingual Data** **27**

3.1 Introduction: Why Multilingual Data? . . . . . 27

3.2 Parallel vs. Comparable Data . . . . . 28

3.2.1 Parallel Corpora . . . . . 29

3.2.2 Comparable Corpora . . . . . 30

3.3 Conclusions . . . . . 31

**4 Multilingual Probabilistic Topic Modeling** **33**

4.1 Introduction . . . . . 33

4.2 A General Framework . . . . . 36

4.2.1 Definitions and Assumptions . . . . . 36

4.3 A More General Framework: Latent Cross-Lingual Concepts (Intermezzo) . . . . . 40

4.4 Bilingual Latent Dirichlet Allocation (BiLDA) . . . . . 41
4.4.1 An Overview of the Model . . . . . 41
4.4.2 Training: Estimating the BiLDA Model . . . . . 45
4.4.3 Output: Per-Document Topic and Per-Topic Word Distributions . . . . . 50
4.4.4 Inference or "What with New Documents?" . . . . . 50
4.5 Evaluation of Multilingual Topic Models . . . . . 51
4.6 A Short Overview of Other Multilingual Probabilistic Topic Models 53
4.7 Conclusions and Future Work . . . . . 56
4.8 Related Publications . . . . . 57

**II Finding Term Translations in Parallel Data 59**

**5 SampLEX: A New Algorithm for Bilingual Lexicon Extraction from Parallel Data 63**

5.1 Introduction . . . . . 63
5.2 Learning Translation Pairs Using Sub-Corpora Sampling . . . . 65
5.2.1 Why Sampling Sub-Corpora? . . . . . 65
5.2.2 Criteria for Extraction of Translation Pairs . . . . . 66
5.2.3 SampLEX: The Algorithm for Lexicon Extraction . . . 68
5.2.4 Properties of the SampLEX Algorithm . . . . . 71
5.3 State-of-the-Art Models for BLE . . . . . 72
5.3.1 IBM Model 1 . . . . . 72
5.3.2 DICE Model . . . . . 72
5.3.3 LLR Model . . . . . 73
5.4 Experimental Setup . . . . . 73
5.5 Experiments, Results and Discussion . . . . . 75

5.5.1 Experiment I: Testing the Quality of the SampLEX Lexicon in Terms of Precision . . . 76
5.5.2 Experiment II: Investigating Indirect Associations . . . 77
5.5.3 Experiment III: Experiments with a Limited Amount of Parallel Data . . . 78
5.5.4 Experiment IV: Investigating Convergence . . . 81
5.6 Conclusions and Future Work . . . 83
5.7 Related Publications . . . 84

**III Modeling Cross-Lingual Semantic Similarity 85**

**6 A Framework for Modeling Semantic Similarity Based on Latent Cross-Lingual Topics 89**

6.1 Introduction . . . 89
6.2 Cross-Lingual Semantic Similarity: An Overview of Distributional Models . . . 91
6.2.1 Definitions . . . 91
6.2.2 Related Work (Shared Cross-Lingual Features) . . . 92
6.2.3 Quick Notes on Terminology . . . 93
6.3 Cross-Lingual Semantic Similarity via Latent Cross-Lingual Topics 94
6.3.1 Conditional Topic Distributions . . . 95
6.3.2 KL Model and JS Model . . . 96
6.3.3 TCos Model . . . 97
6.3.4 BC Model . . . 97
6.3.5 Cue Model . . . 98
6.3.6 TI Model . . . 98
6.3.7 TI+Cue Model . . . 99
6.3.8 Topic Space Pruning . . . 100

6.4 Experimental Setup . . . 101
6.4.1 Evaluation Task: Bilingual Lexicon Extraction . . . 101
6.4.2 Training, Testing and Evaluation . . . 101
6.5 Experiments, Results and Discussion . . . 103
6.5.1 Experiment I: Comparison of All Models . . . 104
6.5.2 Experiment II: Analysis of Topic Space Pruning . . . 107
6.6 Conclusions and Future Work . . . 109
6.7 Related Publications . . . 110

**7 Selecting Highly Confident Translation Pairs 111**

7.1 Introduction . . . 111
7.2 Main Modeling Assumptions . . . 112
7.2.1 Symmetry Assumption . . . 113
7.2.2 One-to-One Constraint . . . 113
7.3 Algorithm for Selecting Highly Confident Pairs . . . 116
7.3.1 One-Vocabulary-Pass . . . 116
7.3.2 The Final Algorithm: SelRel . . . 117
7.4 Experimental Setup . . . 119
7.5 Experiments, Results and Discussion . . . 120
7.5.1 Experiment I: Do Our Modeling Assumptions Help Bilingual Lexicon Extraction? . . . 120
7.5.2 Experiment II: Thresholding and Precision? . . . 121
7.5.3 Experiment III: Building a Highly Confident Lexicon . . . 122
7.6 Conclusions and Future Work . . . 124
7.7 Related Publications . . . 124

**8 Cross-Lingual Similarity of Words as the Similarity of Their Semantic Word Responses 125**

8.1 Introduction . . . 125
8.2 Modeling Cross-Lingual Word Similarity as the Similarity of Semantic Word Responses . . . 127
8.2.1 The Intuition Behind the Approach . . . 127
8.2.2 Modeling Semantic Responses via Cross-Lingual Topics 128
8.2.3 Response-Based Model of Similarity . . . 129
8.3 Experimental Setup . . . 130
8.4 Experiments, Results and Discussion . . . 132
8.5 Conclusions and Future Work . . . 135
8.6 Related Publications . . . 136

**9 Bootstrapping Cross-Lingual Vector Spaces (from Almost Nothing) 137**
9.1 Introduction . . . 137
9.2 Bootstrapping Cross-lingual Vector Spaces: A Complete Overview 139
9.2.1 General Framework for Bootstrapping . . . 139
9.2.2 Initializing Cross-Lingual Vector Spaces . . . 142
9.2.3 Estimating Confidence of New Dimensions . . . 143
9.3 Experimental Setup . . . 144
9.4 Experiments, Results and Discussion . . . 146
9.4.1 Experiment I: Is Initialization Important? . . . 146
9.4.2 Experiment II: Is Confidence Estimation Important? . . 152
9.5 Conclusions and Future Work . . . 153
9.6 Related Publications . . . 154

**10 Modeling Cross-Lingual Semantic Similarity in Context 155**
10.1 Introduction . . . 155
10.2 Related Work . . . 157

10.3 Towards Context-Sensitive Models of Cross-Lingual Semantic Similarity . . . 159
10.3.1 Why Context-Sensitive Models of Cross-Lingual Semantic Similarity? . . . 160
10.3.2 Defining Context . . . 162
10.3.3 Projecting Context into the Latent Semantic Space . . . 163
10.4 Context-Sensitive Models of Similarity via Latent Cross-Lingual Topics . . . 164
10.4.1 DIRECT-FUSION Model . . . 166
10.4.2 SMOOTHED-FUSION Model . . . 167
10.4.3 LATE-FUSION Model . . . 168
10.5 Experimental Setup . . . 168
10.5.1 Evaluation Task: Word Translation in Context . . . 168
10.5.2 Training, Testing and Evaluation . . . 169
10.6 Experiments, Results and Discussion . . . 172
10.6.1 Experiment I: Results on the CWT+JA-BNC Test Set . 173
10.6.2 Experiment II: Results on the CWT+Wiki Test Set . . 173
10.6.3 Experiment III: Analysis of Context Sorting and Pruning 177
10.7 Conclusions and Future Work . . . 178
10.8 Related Publications . . . 180

**IV Cross-Lingual Information Retrieval 181**

**11 Multilingual Topic Models in (Cross-Lingual) Information Retrieval 185**

11.1 Introduction . . . 185
11.2 Related Work . . . 188
11.3 MuPTM-Based CLIR . . . 189
11.3.1 MuPTM-Basic CLIR Model . . . 189

11.3.2 MuPTM-DM CLIR Model . . . 192
11.3.3 MuPTM-SemLex CLIR Model . . . 194
11.4 Experimental Setup . . . 198
11.4.1 Evaluation Tasks: Known-Item Search and Ad-Hoc Search 198
11.4.2 Training Collections . . . 198
11.4.3 Test Collections and Queries . . . 200
11.4.4 Training Setup and Parameters of CLIR Models . . . 203
11.4.5 Evaluation Metrics . . . 204
11.5 Experiments, Results and Discussion . . . 205
11.5.1 A Short Overview . . . 205
11.5.2 Experiment 0: Cross-Lingual Known-Item Search for Wikipedia Articles . . . 206
11.5.3 Experiment I: Comparison of the MuPTM-Basic Model with Baseline Models . . . 208
11.5.4 Experiment II: Comparison of MuPTM-Basic, MuPTM-DM and MuPTM-SemLex . . . 209
11.5.5 Experiment III: Comparison with Monolingual MoPTM-Based Models and MuPTM-Based Models that Use an External Translation Resource . . . 214
11.5.6 Experiment IV: Comparison of MuPTM-SemLex and SemLex-Basic . . . 216
11.5.7 Experiment V: Training with Different Types of Corpora 217
11.5.8 Experiment VI: Perplexity and Retrieval . . . 219
11.6 Conclusions and Future Work . . . 219
11.7 Related Publications . . . 221

**12 MuPTM and (Cross-Lingual) Relevance Modeling 223**

12.1 Introduction . . . 223
12.2 A Short Introduction to Relevance Modeling . . . 225

12.3 Cross-Lingual Estimation of a Relevance Model . . . . . 226
12.3.1 Prior Work . . . . . 226
12.3.2 Approximating a Cross-Lingual Relevance Model . . . . 227
12.3.3 Making the Model Tractable . . . . . 228
12.3.4 Estimating CRM by MuPTM . . . . . 229
12.3.5 Final Retrieval Model . . . . . 230
12.4 Experimental Setup . . . . . 231
12.4.1 Training Data, Test Data and Evaluation Metrics . . . . . 231
12.4.2 Models for Comparison . . . . . 231
12.4.3 Parameters . . . . . 232
12.5 Experiments, Results and Discussion . . . . . 232
12.6 Conclusions and Future Work . . . . . 235
12.7 Related Publications . . . . . 236

**13 Conclusions and Future Perspectives** **237**
13.1 Thesis Summary . . . . . 237
13.2 Contributions . . . . . 241
13.3 Future Work . . . . . 242

**A Appendix - TermWise CAT Tool** **245**
A.1 Why Term&Phrase Memory in a CAT Tool? . . . . . 246
A.2 Knowledge Acquisition . . . . . 247
A.3 Context-Sensitive Database Querying . . . . . 249
A.4 Evaluation . . . . . 250
A.5 Related Publications . . . . . 250

**Bibliography** **251**

**Curriculum Vitae** **273**

**List of Publications** **275**

# List of Figures

1.1 Outline of the thesis. . . . . 9

2.1 Examples of several (a) binomial distributions with different values for $p_s$ and $n_s = 20$, and (b) Beta distributions with varying $\alpha$ and $\beta$. . . . . 16

4.1 An illustrative overview of the key intuitions behind multilingual probabilistic topic modeling. . . . . 39

4.2 Graphical representation of the bilingual LDA (BiLDA) model in plate notation. $R^S$ and $R^T$ denote lengths of the source document and the target document in terms of word tokens for each aligned document pair. . . . . 43

4.3 Polylingual topic model: The generalization of the BiLDA model which operates with $l$ languages, $l \geq 2$. . . . . 45

4.4 Standard monolingual LDA model from Blei et al. [31]. . . . . 45

5.1 Precision and F-1 scores over (a) Dutch-English, and (b) Italian-English parallel corpora of different size (2k, 10k, 50k sentence pairs). . . . . 79

5.2 Precision, recall and F-1 scores for Dutch-English over the sequence of sampling rounds in the first iteration of the SampLEX algorithm. . . . . 81

5.3 Precision, recall and F-1 scores for Italian-English over the sequence of sampling rounds in the first iteration of the SampLEX algorithm. . . . . 81

5.4 Precision, recall and F-1 scores over the first 10 iterations of SampLEX for (a) Dutch-English, and (b) Italian-English. . . . 82

6.1 $Acc_1$ and $Acc_{10}$ scores for KL-MuPTM, JS-MuPTM, TCos-MuPTM, BC-MuPTM, and comparisons with baseline models: KL-MoPTM, JS-MoPTM, TCos-MoPTM, BC-MoPTM, and BaseCos. . . . . 104

6.2 $Acc_1$ and $Acc_{10}$ scores for Cue-MuPTM, TI-MuPTM, TI+Cue-MuPTM, and comparisons with baseline models: Cue-MoPTM, TI-MoPTM, TI+Cue-MoPTM, and BaseCos. . . . . 105

6.3 $Acc_1$, $Acc_{10}$, $MRR$ scores for JS-MuPTM and BC-MuPTM along with their execution times. The horizontal axis is in logarithmic scale. . . . . 108

7.1 An illustrative example depicting the basic advantages of introducing the symmetry assumption and the one-to-one constraint. . . . . 116

7.2 Precision and $F_{0.5}$ scores in relation to threshold values. . . . . 121

9.1 An illustration of the bootstrapping approach to constructing a shared Spanish-English cross-lingual vector space. . . . . 140

9.2 Results with 3 different seeding methods as starting points of the bootstrapping process: (i) identical words only (SEED-ID), (ii) the BC-Topics model (SEED-TB), (iii) the BC-Responses model (SEED-RB). (a) $Acc_1$ and $Acc_{10}$ scores for ES-EN, (b) $Acc_1$ and $Acc_{10}$ scores for IT-EN. . . . . 147

9.3 The number of dimensions in the cross-lingual vector space with the 3 different seeding methods in each iteration for ES-EN and IT-EN. The bootstrapping procedure typically converges after a few iterations. . . . . 148

9.4 Results on the BLE task with SEED-RB when using seed translation pairs of different frequency: (i) high-frequency (HF-SEED), (ii) medium-frequency (MF-SEED), (iii) low-frequency (LF-SEED). . . . . 149

9.5 The effect of learning rate $B$ on bootstrapping. ES-EN. Seed lexicon: SEED-RB with 600 pairs, confidence function: symmetrized M-Best. . . . . 152

10.1 An example of cross-lingual word similarity without and with context. The lists contain English words similar to Italian word *campione* before observing any context and after observing a context word *squadra (team)*. . . . . 160

10.2 An illustrative toy example of the main intuitions in our probabilistic framework for building context-sensitive models of cross-lingual semantic similarity. . . . . 165

10.3 The influence of the size of sorted context on the accuracy of word translation in context. Test dataset is CWT+Wiki. The model is SMOOTHED-FUSION (SF=Cue). . . . . 177

11.1 MuPTM-Basic retrieval model: An illustrative graphical presentation of the basic retrieval model that relies only on the latent layer of cross-lingual topics obtained by a multilingual probabilistic topic model. . . . . 191

11.2 An example of a probabilistic semantic lexicon entry from a Dutch-English lexicon obtained from top $M = 5$ words from the ranked list. The scores on the edges on the left side are unnormalized similarity scores (the higher the score, the higher their semantic similarity). The scores on the edges on the right side (after the thick arrow) present normalized probability scores $P(w_j^T|w_1^S)$ after eq. (11.7) is employed. . . . . 196

11.3 11-pt recall-precision curves for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex for both retrieval directions. Multilingual topic model is BiLDA, $K = 1000$. Training corpus is Wiki+EP. 210

11.4 11-pt recall-precision curves for MuPTM-DM and MuPTM-SemLex for all CLEF test collections. Multilingual topic model is BiLDA. Training corpus is Wiki+EP. . . . . 211

11.5 Comparison of DM-Basic, MuPTM-Basic, and MuPTM-DM as their combination. BiLDA with $K = 1000$. Training corpus is Wiki+EP. . . . . 212

11.6 11-pt recall-precision curves for SemLex-Basic and MuPTM-SemLex. BiLDA. Training corpus is Wiki+EP. . . . . 216

11.7 Comparison of the 11-pt recall-precision curves values for MuPTM-SemLex, where BiLDA was trained on different corpora (EP, Wiki and Wiki+EP). $K = 1000$. . . . . 218

12.1 Queries and relevant documents are random samples from an underlying relevance model $R_Q$ (left). A dependence network for the estimation of the joint probability $P(w^T, q_1^S, \ldots, q_m^S)$ (right): The query words $q_1^S, \ldots, q_m^S$ and the words $w^T$ are sampled independently and identically from a distribution representing the document $d_j^T$. . . . . 226

12.2 11-pt recall-precision curves for all models over all campaigns. The positive synergy between probabilistic topic modeling and relevance modeling is clearly visible in both the monolingual setting and the cross-lingual setting. A similar relative performance is observed in the reverse retrieval direction (Dutch queries, English documents) and in the English monolingual retrieval task. . . . . 234

A.1 TermWise CAT tool: an overview of its architecture. . . . . 246

# List of Tables

4.1 Randomly selected examples of latent cross-lingual topics represented by top 10 words based on their counts after Gibbs sampling. Topics are obtained by BiLDA trained on Wikipedia for various language pairs: French-English (FR-EN), Dutch-English (NL-EN), Italian-English (IT-EN), and Spanish-English (ES-EN). For non-English words we have provided corresponding English translations. $K = 100$ for all models. . . . . 52

5.1 The contingency table for a pair of words $(w_1^S, w_2^T)$. . . . . 73

5.2 Precision and MRR scores for all models trained on the first $300,000$ sentences of Dutch-English Europarl data, and evaluated on the sets of $1,001$ ground truth translation pairs for Dutch-English. . . . . 76

5.3 Precision and MRR scores for all models trained on the first $300,000$ sentences of Italian-English Europarl data, and evaluated on the sets of $1,001$ ground truth translation pairs for Italian-English. All models (including SampLEX) provide translations for all $1,001$ from the ground truth test set. . . . . 76

5.4 Precision and MRR scores on our evaluation set consisting of Italian *-iamo* verbs (present tense, first person plural). . . . . 78

5.5 Precision and MRR scores on Dutch-English for all models trained on the subsets of different sizes (2k, 10k, 50k sentences). . . . . 80

5.6 Precision and MRR scores on Italian-English for all models trained on the subsets of different sizes (2k, 10k, 50k sentences). 80

6.1 Best $Acc_1$ and $Acc_{10}$ scores over all values of $K$ (in parentheses after each result) for all compared models. . . . . 105

6.2 Lists of the top 5 semantically similar words (Italian to English), where the correct translation candidate is not found (column 1), lies hidden lower in the pruned ranked list (2), and is retrieved as the most similar words (3). All three lists are obtained with TI+Cue-MuPTM. . . . . 106

6.3 Topic space pruning: $Acc_1$, $MRR$, and $Acc_{10}$ scores for JS-MuPTM, TCos-MuPTM and BC-MuPTM which rely on word representations by means of conditional topic distributions over different values of pruning parameter $K'$. BiLDA. $K = 2000$. . 108

7.1 $Acc_1$ scores for 2 language pairs with our 4 BLE algorithms. . . . 121

7.2 A comparison of different precision-oriented bilingual lexicons for Italian-English and Dutch-English in terms of the number of correct translation pairs, precision and $F_{0.5}$ scores. . . . . 123

8.1 An example of top 10 semantic word responses and the final response-based similarity (last column) for a selection of Spanish and English words. The responses are estimated from Spanish-English Wikipedia data by BiLDA. . . . . 130

8.2 Results on the BLE task. Language pairs are Italian-English and Spanish-English. . . . . 132

8.3 Results on the BLE task for Dutch-English, with different corpora used for the estimation of semantic word responses. . . . . 132

8.4 Example lists of top 10 semantically similar words across all three language pairs according to our BC-Responses similarity model, where the correct translation word is: (column 1) found as the most similar word, (2) contained lower in the list, and (3) not found in the top 10 words in the ranked list. . . . . 133

8.5 Example translation pairs found by BC-Responses, but missed by the other three compared models of similarity. . . . . 134

9.1 ES-EN: Results with different sizes of the seed lexicon. The number in the parentheses denotes the number of dimensions in the cross-lingual space after the bootstrapping procedure converges. The seeding method is SEED-RB. . . . . 151

9.2 IT-EN: Results with different sizes of the seed lexicon. The number in the parentheses denotes the number of dimensions in the cross-lingual space after the bootstrapping procedure converges. The seeding method is SEED-RB. . . . . 151

10.1 Sets of 15 ambiguous words in Spanish, Italian and Dutch from the CWT+Wiki dataset accompanied by the sets of their respective possible senses/translations in English. . . . . 170

10.2 Example sentences from our CWT+JA-BNC and CWT+Wiki evaluation datasets with the corresponding correct word translations from the ground truth. . . . . 171

10.3 Results on the CWT+JA-BNC test dataset. Training corpus is Wiki. Translation direction is EN-ES/IT. . . . . 174

10.4 Results on the CWT+JA-BNC test dataset displaying the difference in results when training on Wiki and Wiki+EP. Translation direction is EN-NL. . . . . 174

10.5 Results on the CWT+Wiki test dataset. Training corpus is Wiki. Translation direction is ES/IT-EN. . . . . 175

10.6 Results on the CWT+Wiki test dataset displaying the difference in results when training on Wiki and Wiki+EP. Translation direction is NL-EN. . . . . 175

10.7 Results for different sizes of sorted context sets. Test dataset is CTW+Wiki. The model is SMOOTHED-FUSION (SF=Cue). 178

11.1 Lists of the top 10 translation candidates (Dutch to English), where the correct translation is not found (column 1), lies hidden lower in the list (2), and is retrieved as the first candidate (3). Obtained with the TI+Cue method of cross-lingual semantic similarity. . . . . 195

11.2 Statistics of the CLEF 2001-2003 CLIR test collections. . . . . 203

11.3 Statistics of used queries (CLEF test collections). . . . . 203

11.4 $Acc_1$ and $Acc_5$ scores in both search directions for DM-Basic and SemLex-Basic in the cross-lingual known-item search for Wikipedia articles. . . . . 206

11.5 $Acc_1$ and $Acc_5$ scores for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex in the cross-lingual known-item search. EN queries, NL target articles. . . . . 207

11.6 $Acc_1$ and $Acc_5$ scores for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex in the cross-lingual known-item search. NL queries, EN target articles. . . . . 207

11.7 MAP scores on all CLEF test collections for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex, where BiLDA was trained with different number of topics (400, 1000, 2200). Training corpus is Wiki+EP. . . . . 210

11.8 MAP scores on all CLEF test collections for MoPTM-Basic, MoPTM-DM, GT+MoPTM-Basic and GT+MoPTM-DM. Standard monolingual LDA trained on monolingual English and Dutch data. Wiki+EP. $K = 1000$. . . . . 214

11.9 MAP scores on all CLEF test collections for MuPTM-DM and MuPTM-SemLex, where BiLDA was trained on different corpora (EP, Wiki, and Wiki+EP). $K = 1000$. . . . . 217

11.10 Perplexity scores after the inference of the BiLDA model (trained on the Wiki+EP corpus) on the CLEF test collections. . . . . 219

12.1 MAP scores on the CLEF monolingual and cross-lingual retrieval task with English (and Dutch) queries and Dutch document collection. All relative performances are given with respect to the baseline MRM+DM model performance. Each model is also assigned a unique symbol. The symbols indicate statistically significant differences between the MAP scores in each campaign of every two models to which these symbols are assigned. We use the two-tailed t-test ($p < 0.05$). . . . . 233

A.1 Example output of the SampLEX algorithm operating with $N$-gram candidate terms. Translation direction: French to Dutch. 249

# List of Algorithms


4.1 Generative story for BiLDA . . . . . 43
4.2 Gibbs sampling: A general overview . . . . . 47
4.3 Gibbs sampling for BiLDA: An overview . . . . . 47
5.1 One sampling round with fixed $\omega$ . . . . . 69
5.2 SampLEX algorithm . . . . . 70
6.1 Topic space pruning . . . . . 101
7.1 Symmetrizing re-ranking . . . . . 114
7.2 One-Vocabulary-Pass . . . . . 117
7.3 SelRel algorithm . . . . . 118
9.1 Bootstrapping a cross-lingual vector space . . . . . 141
11.1 MuPTM-Basic retrieval model . . . . . 192
11.2 MuPTM-SemLex retrieval model . . . . . 199
12.1 Relevance modeling with MuPTM . . . . . 230

# 1 Introduction

*"Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."*

— Lewis Carroll

New information technologies have enabled an extremely powerful and fast flow of information and have made our interconnected world a true *global village*. The ideas behind the semantic Web, blogosphere and social media, and a remarkable growth in the numbers of Web users around the world have sparked an evolutionary process which continues to change the face and surface of today's *world of information*. Following the ongoing growth of the World Wide Web and its omnipresence in today's increasingly connected world, users tend to abandon English as the *lingua franca* of the global network, since more and more content becomes available in their native languages. In short, more users generate and browse through more data in more different languages. In other words, the users have simultaneously generated a huge volume of multilingual text resources. Available text data, originally dominated by the use of English as the omnipresent language of the Web, now occurs in a plethora of different languages and dialects.[1] The world has become a *data-driven multilingual*

[1] It is difficult to determine the exact number of languages in the world, but the estimations vary between 6,000 and 7,000 official languages. Source `http://www.ethnologue.com` lists 6,912 main languages in its language index. Even more remarkable is the number of different dialects and unofficial languages which, according to Ethnologue, rises up to 39,491. Furthermore, a recent trend on the Web is the presence of different community languages which are all idioms of the same language (e.g., social media consumers, e-commerce retailers, scientists or legislation typically differ in their idiomatic usage of the same language).

*environment* overloaded with multilingual information. Considering the rapidly growing amount of multilingual data and the number of unofficial and official languages in the world, there is a pressing need to provide tools that are able to cross the language barriers in a cheap and effective manner. In other words, the aim is to induce knowledge from the user-generated multilingual text resources and effectively accomplish multilingual text processing automatically or with minimum human intervention. There is another need to facilitate navigation by users through the huge and constantly thriving world of multilingual information. Following these observations, we have detected these key requirements for the automatic tools for multilingual text processing:

R1. Considering the large number of languages (and, consequently, language pairs), we need to represent multilingual text content, that is, words, phrases, documents written in different languages in a structured and coherent way, regardless of their actual language. In short, we require *language-independent* and *language pair independent representations* of multilingual text content.

R2. The users should be able to understand the content which is not given in their native languages. A fundamental cross-lingual problem is the problem of *meaning* and cross-lingual *semantics*, which implies the problem of *translation*. We require widely applicable tools that are able to automatically detect a similar meaning of text units across a wide variety of language pairs and induce translational resources (e.g., bilingual lexicons) directly from the data itself.

R3. The users should be able to acquire knowledge and satisfy their information need from the relevant content which is not always given in their native languages. Therefore, we require tools which deal with another fundamental problem of *information retrieval* (monolingually and across languages). These tools should again be applicable to a wide variety of languages and language pairs.

R4. Besides the requirement of being *applicable to a wide spectrum of languages, idioms and domains*, the tools have to be *cheap*, *data-driven* and *effective*.

This dissertation revolves around this set of requirements. In order to provide solutions which address these requirements and to spark more research interest in the relevant research domains, we had to hunt high and low for the possible answers. The final result of this "hunt" is this thesis text. It stands at the crossroads of three major fields within the immense and spectacular universe

called *artificial intelligence* (AI). It draws its inspiration from and explores the depths of (1) *data mining*, (2) *computational linguistics* (CL) and *natural language processing* (NLP), and (3) *information retrieval* (IR). Some readers might even say that, in some parts, it even trespasses into the domains of cognitive science and psycholinguistics.

**Artificial Intelligence.** It is technology and a branch of computer science that studies and aims to develop intelligent machines and software. It may be ambitiously defined as a part of human technological and scientific activity that aims to build systems that think and do as humans. A more practical definition states that AI aims to build systems that perform tasks that simulate intelligent behavior. We do not want to further raise a discussion concerning the actual definition of intelligence and intelligent behavior within this particular context [297, 265]. In this thesis, we explore tools which aim to address the problems of meaning, translation and information retrieval, where the goal is to alleviate these tasks and help humans deal with the multilingual information overload.

**Data Mining.** It is an interdisciplinary branch of computer science and denotes the process of discovering patterns in large data sets [86]. In this thesis we explore a new modeling approach called multilingual probabilistic topic modeling which is able to induce language-independent and language pair independent representations of words and documents automatically from raw multilingual text data.

**Natural Language Processing.** It is a field of research between computer science and linguistics which aspires towards the automated analysis, representation, transformation and generation of natural language texts by means of computer algorithms [186, 141]. With the huge increase of multilingual text content, *multilingual natural language processing* which aspires towards solving cross-lingual tasks has gained much interest recently. A large part of this thesis is situated within this sub-field of NLP, with a special focus on (cross-lingual) *semantics*. We explore different algorithms for inducing semantic knowledge from multilingual text data analysis, that is, we propose and investigate models of meaning and similarity across languages which are able to deal with the multilingual text data.

**Information Retrieval.** Information retrieval is the activity of obtaining information resources relevant to a user's information need from a large collection of information resources (e.g., documents in a document collection) [185]. Text-based information retrieval systems often rely on NLP techniques and these two areas of research are often intertwined. With the growing volume of multilingual data and the presence of very specific user needs, modern information retrieval systems widely require the support of *cross-lingual information retrieval*, that is, finding documents in a language different from the user's query. In this thesis,

we explore cross-lingual information retrieval models which should be widely applicable to a variety of language pairs as they do not utilize any language pair dependent knowledge.

## 1.1 Motivation and Goals

An important property of multilingual text collections or corpora addresses their *parallelity* or *comparability*. A parallel corpus is a dataset provided in more than one language, where each document has an exact counterpart in every other language (i.e., an exact translation). On the other hand, comparable corpora do not possess direct exact translations, but there exists a similarity (at least to a certain extent) between the text collections given in different languages. Documents in comparable corpora typically address a portion of similar themes or subjects and could be observed as theme-aligned corpora. Different corpora types typically require different tools that are able to induce knowledge from the data.

Going back to requirements R1-R4 and knowing the main properties of different multilingual text collections, the goal of this thesis is to provide solutions and answers to the questions raised from these requirements with respect to a variety of multilingual text corpora. In short, we formulate the following research questions which have mainly driven the research conducted in this thesis:

(RQ1) How to represent multilingual text content, that is, words, phrases and documents written in different languages in a structured and coherent way, regardless of their actual language? How to induce these language independent and language pair text representations?

(RQ2) Is it possible to automatically build statistical data-driven models of cross-lingual semantic similarity (i.e., addressing the *problem of meaning*) for a variety of language pairs, without any prior knowledge about language pairs, and without any high-quality external resource (e.g., a hand-crafted domain-specific bilingual lexicon)? Is it possible to automatically induce translational resources such as bilingual lexicons (i.e., addressing the *problem of translation*) in the same manner for a variety of language pairs?

(RQ3) How to deal with uncertainty in data and extract only highly reliable translation pairs from such noisy environments?

(RQ4) Is it possible to construct robust and cheap statistical algorithms for cross-lingual information retrieval again without any external translation

resources for a variety of language pairs (i.e., addressing the problem of *information retrieval*)? Since the cross-lingual setting is a more general setting, is it possible to transfer the same modeling principles to build more robust monolingual information retrieval models?

(RQ5) Do we require different algorithms to deal with different types of multilingual text corpora? In other words, do we have to change the algorithmic approach when moving from the multilingual setting with parallel data to the setting with only comparable data at our disposal?

In order to address requirement R1 and research question RQ1, we systematically study the framework of multilingual probabilistic topic modeling. The framework enables language-independent and language pair independent representations of words and documents by means of the shared set of latent cross-lingual topics/concepts which are induced from multilingual text data directly. We also investigate the utility of these representations in two major cross-lingual tasks: (1) modeling cross-lingual semantic similarity and the related task of bilingual lexicon extraction (research questions RQ2 and RQ3), (2) modeling cross-lingual information search and retrieval (research question RQ4).

Furthermore, in order to automatically build automatic translation resources such as bilingual lexicons (requirement R2 and research question RQ2), we investigate and propose different algorithms for parallel and comparable data (addressing research question RQ5). Since comparable datasets are typically more abundant, address more domains and topics, and are much cheaper to acquire (therefore, they are better aligned with requirement R4), we propose and investigate language pair independent and cheap models of cross-lingual information retrieval (research question RQ4) which can be built solely on the basis of comparable training data.

The work in this thesis is situated within the broad framework of *probabilistic modeling*. Probabilistic modeling is a mainstay of modern artificial intelligence research, providing essential tools for analyzing and inducing knowledge from the vast amount of available data [27]. Besides being interpretable and intuitive, the probabilistic modeling approach aims to faithfully represent uncertainty, parameters and noise in observed data. It aspires towards building automated and adaptive models which exhibit robustness and scale well to large data collections [27, 217]. It also allows for including complementary pieces of evidence and sources of information during modeling, which altogether makes probabilistic modeling an extremely versatile modeling framework.

It is also worth mentioning that the algorithms and tools described in this thesis

are *unsupervised*. It means that no human annotation is necessary prior to the learning process: all required information is contained in the raw dataset, that is, all proposed models are *corpus-based* and *data-driven*. Such approaches are again aligned with requirement R4 and a desire to provide algorithms and tools which will be as generally applicable as possible.

A large portion of the work reported in this dissertation is closely aligned with and contributes to the practical requirements of the research projects that provided funding for this research.

## 1.2 Contributions

Following the motivation from sect. 1.1, the following major contributions will be presented in this thesis, listed in "chronological order" (i.e., as they appear in the thesis text):

I. The first full systematic and comprehensible overview of the recently developed multilingual probabilistic topic modeling (MuPTM) framework, with its theoretical background, definitions, modeling assumptions, methodology and applications described all the way up from the conceptual level and the modeling level down to the mathematical foundations. We demonstrate how to utilize MuPTM to induce the set of latent cross-lingual topics from comparable data, and their ability to provide language-independent and language pair independent representations of words and documents. This contribution is discussed in **part I** of this thesis.

II. A new language pair independent approach to designing statistical models that relies on the paradigm of data reduction, sub-corpora sampling and utilizing low-frequency events/words (which are often disregarded in statistical models as statistically insignificant and unreliable). The approach is applied in the task of bilingual lexicon extraction (BLE) from parallel corpora. We present a new algorithm for bilingual lexicon extraction from parallel data that outperforms the established state-of-the-art BLE models. This contribution is discussed in **part II**.

III. A new language pair independent statistical framework for modeling cross-lingual semantic similarity (out of context and in context) from comparable corpora which relies on the notion of latent cross-lingual topics/concepts. As a follow-up contribution, we demonstrate how to induce bilingual lexicons from comparable data (as opposed to inducing them from parallel data in part II). This contribution is discussed in **part III**.

IV. A new language pair independent approach to cross-lingual information retrieval and the construction of a new MuPTM-based probabilistic framework for cross-lingual information retrieval which allows for embedding many additional evidences in building new retrieval models without any additional external resources such as machine-readable translation dictionaries or parallel corpora. This contribution is discussed in **part IV**.

Other, more focused contributions are subsumed by these major broader contributions. For instance, in part III, as these more focused contributions, we propose an algorithm for selecting only highly reliable translation pairs from comparable corpora, a bootstrapping approach to modeling cross-lingual semantic similarity or context-sensitive models of similarity, etc. These more focused contributions will be gradually introduced in each part of the thesis.

## 1.3 Structure of the Thesis

The rest of the thesis is organized as follows. Chapter 2 gives an extended introduction to the key background concepts needed to understand the rest of the thesis with an emphasis on the fundamentals of probability theory and probabilistic modeling. The reader familiar with these concepts may safely skip this chapter.

Following that, in part I we tackle the problem of multilingual text mining. In particular, we show how to perform the exploitation of multilingual text resources by means of the recently developed multilingual probabilistic topic modeling framework. First, we present and classify multilingual text resources in more detail in chapter 3. We list main properties of parallel and comparable corpora along with their example corpora where the emphasis is on parallel and comparable corpora utilized in this work. Chapter 4 presents a full overview of the multilingual probabilistic topic modeling framework. We show how to train, infer, use and evaluate multilingual probabilistic topic models. Most importantly, we present how to obtain language-independent and language pair independent representations of words and documents by means of multilingual probabilistic topic models (research question RQ1) as these representations are extensively used in the other parts of the thesis.

In part II, we present how to automatically find (highly reliable) term translations in parallel data and induce bilingual lexicons solely on the basis of the given parallel data (research questions RQ2, RQ3 and RQ5). Our new algorithm for bilingual lexicon extraction which relies on a new idea of data reduction and sub-corpora sampling is proposed, evaluated and compared against other state-of-the-art BLE models in chapter 5. Furthermore, we employ

the same modeling principle to perform term alignment and build a full-fledged CAT (computer-assisted translation) tool where this algorithm is run in the background as one of the modules. A brief overview of this tool is given in appendix A.

Part III deals with modeling semantic similarity and inducing bilingual lexicons from comparable corpora (research questions RQ2, RQ3 and RQ5). The proposed statistical framework relies on the knowledge of latent cross-lingual topics from part I. The proposed MuPTM-based framework for modeling cross-lingual similarity is the first known work that combines distributional semantics and multilingual probabilistic topic modeling. We make a series of contributions in part II. These contributions are presented in standalone chapters. First, in chapter 6 we introduce the framework for modeling semantic similarity across languages and automatically extracting bilingual lexicons based on latent cross-lingual topics/concepts. We present a set of models which operate directly in the latent cross-lingual semantic space spanned by these latent topics. Following that, in chapter 7, we present an approach to selecting only highly confident translation pairs from the lists of semantically similar words (tackling research question RQ3). The precision-oriented selection algorithm is extremely important in the noisy setting dealing with comparable data. We introduce a new more robust cross-lingual semantic space spanned by all vocabulary words in both languages in chapter 8. In chapter 9, we propose, present and analyze a new framework for bootstrapping cross-lingual vector spaces from comparable corpora. Finally, in chapter 10, we demonstrate how to build context-sensitive models of semantic similarity within the same statistical framework for modeling cross-lingual semantic similarity.

Part IV proposes a new probabilistic language pair independent framework for (cross-lingual) information retrieval based on the knowledge of latent cross-lingual topics (which may be induced from comparable corpora) and provides extensive evaluations and comparisons of different retrieval models built within this new retrieval framework (research questions RQ4 and RQ5). Part IV is divided into two chapters. In chapter 11, we present the key intuitions behind our retrieval framework and introduce the first set of MuPTM-based retrieval models. Chapter 12 shows how to combine the power of multilingual probabilistic topic modeling with the power of relevance modeling and how to build new more robust monolingual and cross-lingual retrieval models.

Finally, conclusions, contributions and future work perspectives based on the work conducted in this thesis are summarized in chapter 13.

The material presented in this thesis may be observed from multiple angles: (i) Part II and part III show how to induce translational knowledge in two different multilingual settings: while part II presents how to automatically

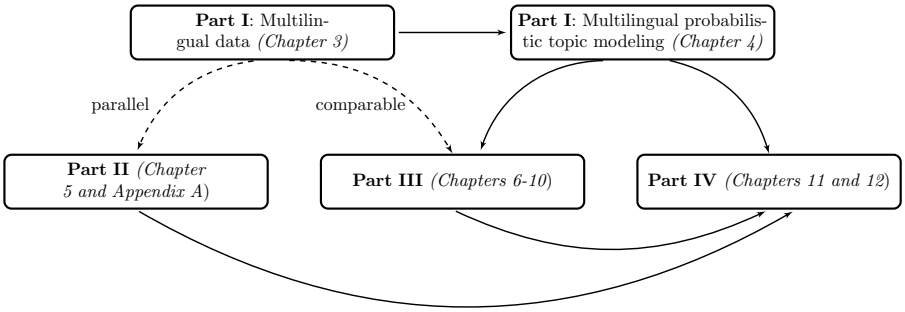



Figure 1.1: Outline of the thesis.

extract bilingual lexicons from parallel data, part III presents how to extract such lexicons from comparable non-parallel data; (ii) Part III *per se* is a detailed overview of cross-lingual distributional models of semantic similarity from comparable data; (iii) Part I *per se* provides a complete and systematic overview of the multilingual probabilistic topic modeling framework; (iv) Part IV tackles the fundamental task of cross-lingual information retrieval, provides invaluable insights and an introduction to this research domain; it also shows how to utilize bilingual lexicons induced either from parallel data (part II) or comparable data (part III) as sources of knowledge in the retrieval process; (v) Part III and part IV in combination show how to utilize multilingual probabilistic topic models (from part I) trained on comparable corpora in two major cross-lingual tasks: modeling cross-lingual semantic similarity and cross-lingual information retrieval respectively. This rather complex combination of viewpoints induced the structure of the thesis and influenced its final presentation. The outline of the thesis is summarized in fig. 1.1.

# 2 Fundamentals

*If I had only one day left to live, I would spend it in my statistics class: it would seem so much longer.*

— Anonymous

## 2.1 Introduction

This chapter contains brief reviews of basic modeling concepts which serve as a fundamental theoretical background for the remainder of this text. The reader must be aware that these fundamentals have not been covered in its entirety, but rather presented in a shortened and concise manner, accompanied with references for further reading. Moreover, the focus of this chapter is only on fundamentals that the author deems necessary for a deeper understanding of the upcoming parts of the thesis text. The reader already familiar with these elementary concepts may safely skip the parts of this chapter discussing them.

As already hinted in chapter 1, the large body of work within this thesis is situated within the probabilistic framework and relies on *Bayesian* modeling. Therefore, in sect. 2.2, we present a short overview of basics of probability theory and Bayesian modeling. Terms and concepts such as probability distributions, joint and conditional probabilities, prior and posterior probabilities, conjugate priors are explained here. Moreover, we present a short overview of the well-known probability distributions (e.g., multinomial and Dirichlet distributions) which are extensively used throughout this thesis. Following that, we briefly

discuss the Bayesian modeling framework, and a difference between hidden (or latent) and observed variables.

Sect. 2.3 presents the basics of statistical language modeling, with an emphasis on probabilistic language models and their utility in information retrieval.

Finally, sect. 2.4 explains some basic text preprocessing steps (e.g., tokenization, stop words removal, part-of-speech tagging) which are further utilized in this thesis for corpora preprocessing.

## 2.2 A (Too) Short Introduction to Probability Theory and Bayesian Modeling

Imagine an event happening with several different possible outcomes (e.g., a flip of a coin or a roll of a die). We would like to know the probability that a coin will land heads. But, what is actually *probability*? What does the phrase "a coin will land heads" actually mean? A *frequentist* interpretation of probability represents probabilities as long run frequencies of events (e.g., if we flip a coin many times, we expect it to land heads about half of the times). On the other hand, the *Bayesian* interpretation of probability models uncertainty about events that do not have long term frequencies (e.g., we *believe* that the coin is equally likely to land tails or heads in the next flip) [137, 217]. In short, in this interpretation, probability is used to quantify our uncertainty or our belief about something. However, regardless of the actual interpretation, the rules of probability theory remain the same.

### 2.2.1 Basic Concepts of Probability Theory

More formally, we may define several basic elements to describe probability:

(i) *Sample space* or the "universe" $\mathcal{W}$ denotes the set of all outcomes of a random experiment. For a single coin flip $\mathcal{W} = \{He, Ta\}$, where $He$ denotes heads and $Ta$ denotes tails after the flip.

(ii) *Event space* $\mathcal{X}$ denotes the set whose elements $X \in \mathcal{X}$ (called *events*) are subsets of $\mathcal{W}$ (i.e., $X \subseteq \mathcal{W}$ is a collection of possible outcomes of an experiment). For a single coin flip $\mathcal{X} = \{\emptyset, \{He\}, \{Ta\}, \{He, Ta\}\}$.

(iii) *Probability measure* denotes a function $P : \mathcal{X} \to \mathbb{R}$ that satisfies the following properties: (1) $P(X) \geq 0$ for all $X \in \mathcal{X}$; (2) $P(\mathcal{W}) = 1$; (3) Any countable sequence of disjoint events $X_1, X_2, \ldots$ (i.e., $X_i \cap X_j = \emptyset$ whenever $i \neq j$) satisfies

$P(\cup_i X_i) = \sum_i P(X_i)$. Again, for a single coin flip, these rules imply $P(\emptyset) = 0$, $P(\{He, Ta\}) = 1$, and $P(\{He\}) + P(\{Ta\}) = 1$.

**Random Variable.** Each partition of the sample space $\mathcal{W}$ is represented by a stochastic or a so-called *random variable* $X$. Formally, a random variable is a function $X : \mathcal{W} \to \mathbb{R}$. The probability of each possible outcome $x$ is the chance that $x$ will happen and is notated as $P(X = x)$. The *probability distribution* $P(X)$ is the distribution of the probabilities over all possible outcomes of $X$. If a random variable can take only possible values/outcomes from a discrete (but finite or at most countably infinite) set of values, it is a *discrete random variable*. On the other hand, if it takes on values that vary continuously within one or more real intervals, it is known as a *continuous random variable*. As a result, the continuous random variable has an uncountably infinite number of possible values, all of which have probability 0, though ranges of such values can have non-zero probability.

**Joint Probability.** Multiple overlapping partitions of the same sample space may be defined, each labeled with a different random variable. The combined probability of multiple variables defined over the same space $\mathcal{W}$ is called the *joint probability*. For two random variables $X$ and $Y$, we denote their joint probability as $P(X, Y)$. This is actually the joint distribution over all possible outcomes for $X$ and $Y$ happening together. The probability $P(X, Y)$ is computed as $P(X \cap Y)$. The probability of a union of two events $X$ and $Y$ is computed as $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$. In case when two events do not overlap (i.e., they are *disjoint*), that probability of at least one of them happening becomes $P(X \cup Y) = P(X) + P(Y)$.

**Conditional Probability and Independence.** If $Y$ is an event with non-zero probability, the *conditional probability* of any event $X$ given $Y$ denoted by $P(X|Y)$ is defined as $P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$. In other words, $P(X|Y)$ is the probability measure of the event $X$ after *observing* the occurrence of the event $Y$. The conditional probability is also called the *posterior* probability since it is computed after the event $Y$ has been observed. It is said that two events $X$ and $Y$ are *independent* if and only if it holds that $P(X \cap Y) = P(X) \cdot P(Y)$ (or equivalently $P(X|Y) = P(X)$). The condition of independence is equivalent to saying that observing $Y$ does not have any effect on the probability of $X$. If $X$ and $Y$ are independent, it consequently holds $P(X, Y) = P(X \cap Y) = P(X|Y) \cdot P(Y) = P(X) \cdot P(Y)$.

**Marginal Probability.** Calculating the probability of only one variable in a joint probability is called *marginalizing* the variable. Given a joint distribution of two events $P(X, Y)$, the marginal probability $P(X)$ is computed by summing the probability over all possible states of $Y$: $P(X) = \sum_Y P(X|Y) \cdot P(Y)$. Here, $P(X)$ is often called the *prior* or *prior distribution*, as it reflects the probability

of the event $X$ in advance, that is, before anything is known about the outcome of the event $Y$.

**Bayes Rule.** An essential rule which will be used throughout the thesis is the so-called Bayes rule or Bayes theorem. The rule describes the relationship between prior and posterior probabilities and is defined as follows:

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \tag{2.1}$$

The rule is simply derived by combining the definition of conditional probability with the product and sum rules: $P(X \cap Y) = P(X|Y) \cdot P(Y) = P(Y|X) \cdot P(X)$. This pivotal rule demonstrates that the posterior probability of $Y$ given $X$ may be expressed in function of the posterior probability of $X$ given $Y$ and the two prior probabilities $P(X)$ and $P(Y)$.

**Pointwise Mutual Information and Mutual Information.** The *pointwise mutual information* (PMI) of a pair of outcomes $x$ and $y$ belonging to discrete random variables $X$ and $Y$ quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence. More formally:

$$\text{PMI}(x;y) = \log \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)} \tag{2.2}$$

$$= \log \frac{P(X=x|Y=y)}{P(X=x)} = \log \frac{P(Y=y|X=x)}{P(Y=y)} \tag{2.3}$$

The mutual information of two random variables is a measure of the mutual dependence of the two random variables. It is actually the expected value of PMI over all possible outcomes of the two variables.

### 2.2.2 Important Probability Distributions

In order to specify the probability measures used when dealing with random variables of any of the two types (i.e., discrete or continuous), it is convenient to specify alternative functions (the so-called *probability functions*) from which the probability measure immediately follows. In other words, a probability function is a mathematical expression which associates a numeric value with each possible outcome of a random variable. We may write $f(x) = P(X = x)$, and $f(x;y) = P(X = x|y)$. There are two distinct variants of probability functions. If we deal with a discrete random variable, we talk about a *probability mass function (pmf)*, while we refer to it as *probability density function (pdf)* in case of a continuous variable.

Several families of discrete and continuous probability functions are well known in the literature. If the probability function of some variable $X$ has a distribution of type $PF$, we say that $X$ is distributed following $PF$, and formally write $X \sim PF$. We review a subset of well known discrete and continuous probability distributions which are utilized further in this thesis.

**Discrete Probability Distributions**. The domain of a *pmf* is always a finite or a countably infinite discrete set of values. The size of the set is known as the *dimension of the distribution*. It holds: $f(x) \in [0, 1]$ and $\sum_x f(x) = 1$.

*Bernoulli Distribution.* It is a two-dimensional distribution which describes the possibility of a "success" in only one trial (e.g., only one coin toss, where "success" might be landing heads). The probability of success is given by parameter $p_s$. We may write $X \sim Bernoulli(p_s)$ and easily compute the following:

$$f(x; p_s) = \begin{cases} p_s & \text{if } x \text{ is a success} \\ 1 - p_s & \text{if } x \text{ is a failure.} \end{cases} \tag{2.4}$$

*Binomial Distribution.* Instead of only one trial as with the Bernoulli distribution, now imagine $n_s$ consecutive trial runs, each again indicated by the success-parameter $p_s$. If $X \sim Binomial(x; n_s, p_s)$, then $P(X = x|n_s, p_s)$ denotes the probability that $x$ out of $n_s$ trial runs have succeeded. We may then compute the probability function as follows:

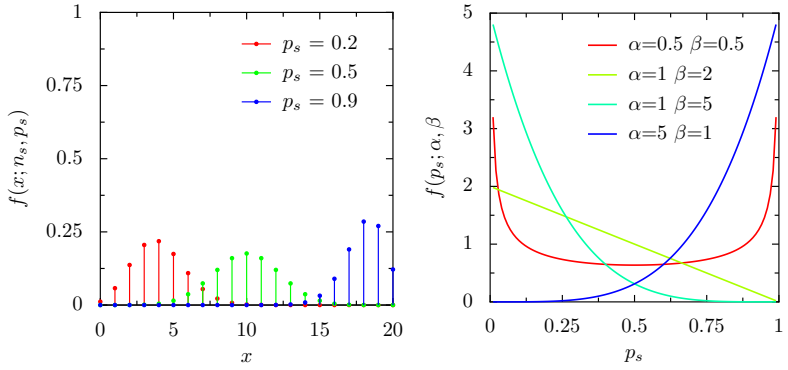
$$f(x; n_s, p_s) = \binom{n_s}{x} \cdot p_s^x \cdot (1 - p_s)^{n_s - x} = \frac{n_s!}{x!(n_s - x)!} \cdot p_s^x \cdot (1 - p_s)^{n_s - x} \tag{2.5}$$

$\binom{n_s}{x} = \frac{n_s!}{x!(n_s - x)!}$ is called a *binomial coefficient*. A few examples of binomial distributions are displayed in fig. 2.1a. Note that the Bernoulli distribution is only a special case of the binomial distribution with only one trial.

*Multinomial Distribution.* The binomial distribution can be used to model the outcomes of coin flips, where only two outcomes are possible: *success* and *failure*. But how to model the outcomes of rolling a $N$-sided die? In other words, we need to model the probability of $n_s$ trials (e.g., $n_s$ rolls of a die) as with the Binomial distribution where one of $N$ outcomes is chosen in each trial, each with probability $p_i$. In this case, a discrete random variable $X$ follows a *multinomial distribution*. The probability that out of $N$ categories in $n_s$ trials, each category $i$ has been chosen $a_i$ times is calculated as follows:

$$f(a_1, \ldots, a_N; n_s, p_1, \ldots, p_N) = \frac{n_s!}{a_1! \cdots a_N!} p_1^{a_1} \cdots p_N^{a_N} \tag{2.6}$$

We may write $X \sim Multinomial(n_s, \mathbf{p})$, where $\mathbf{p}$ is the vector $[p_1, \ldots, p_N]$. The number $\frac{n_s!}{a_1! \cdots a_N!}$ is the *multinomial coefficient*, that is, the number of ways to divide a set of size $n_s = \sum_{i=1}^{N} a_i$ into subsets with sizes $a_1$ up to $a_N$.




(a) Binomial (b) Beta

Figure 2.1: Examples of several (a) binomial distributions with different values for $p_s$ and $n_s = 20$, and (b) Beta distributions with varying $\alpha$ and $\beta$.

*Poisson Distribution.* Another probability distribution expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event. Suppose that someone typically receives 6 e-mails per day on average. However, there will be a certain fluctuation as that person might receive more mails some days, and less some other days, and no mails at all some really bad days. Given only the average rate of mails per day, and assuming that the process is random, the Poisson distribution tells exactly how likely it is that the count of e-mails will be $x$ per day. In other words, the Poisson distribution predicts the degree of spread around a known average rate of occurrence. The $pmf$ of the Poisson distribution for $x = 0, 1, \ldots$ is given as:

$$f(x; \lambda_p) = \frac{\lambda_p^x e^{-\lambda_p}}{x!} \tag{2.7}$$

Here, $\lambda_p$ is the Poisson parameter, and $e$ is the base of the natural logarithm. We may write $X \sim Poisson(\lambda_p)$.

**Continuous Probability Distributions.** A continuous probability density function may have a limited or an unlimited domain. In both cases, variable $X$ has an uncountably infinite number of possible values or outcomes. Since the total probability which is calculated as $\int_x f(x)dx$ has to sum up to 1, this means that the probability of $X \equiv x$ is infinitesimally small. Therefore, a *pdf* only has a meaning over *intervals*: $P(X > A \;\&\; X < B) = \int_A^B f(x)dx$.

*Beta Distribution.* It is closely related to the discrete binomial distribution. In fact, it specifies the opposite. The Beta distribution with parameters $\alpha$ and $\beta$ specifies the probability of the success parameter $p_s$, given the fact that there

have been $\alpha - 1$ successes and $\beta - 1$ failures. The difference here is that $\alpha$ and $\beta$ do not have to be integers, as the Beta distribution is defined for any $\alpha, \beta > 0$ and has support over the interval $[0, 1]$. The distribution is defined as follows:

$$f(p_s; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} p_s^{\alpha-1}(1 - p_s)^{\beta-1}, \tag{2.8}$$

where $B(\alpha, \beta)$ is the Beta function, defined as $\int_0^1 x^{\alpha-1}(1 - x)^{\beta-1}dx$. A few examples of Beta distributions are displayed in fig. 2.1b.

*Dirichlet Distribution.* Another important distribution is the Dirichlet distribution, which stands in an analogous relation with the multinomial distribution as the Beta distribution with the binomial distribution. The Dirichlet distribution is a multivariate generalization of the Beta distribution. Let us have $N$ categories and $\gamma_i$ events assigned with the category $i$, $i = 1, \ldots, N$. The distribution over the $N$-dimensional vector of success parameters $\mathbf{p}$ is then:

$$f(p_1, \ldots, p_N) = \frac{1}{B(\mathbf{p})} \prod_{i=1}^{N} \gamma_i^{p_i - 1}, \tag{2.9}$$

The Beta function $B(\mathbf{p})$ is defined as:

$$B(\mathbf{p}) = B(p_1, \ldots, p_N) = \frac{\prod_{i=1}^{N} \Gamma(p_i)}{\Gamma\left(\sum_{i=1}^{N} p_i\right)} \tag{2.10}$$

The Gamma function $\Gamma(x)$ is defined as:

$$\Gamma(x) = \int_0^\infty t^{(x-1)} e^{(-t)} dt. \tag{2.11}$$

where $t$ is some auxiliary variable.

A difference between two probability distributions defined over the same set may be computed using various measures of similarity or divergence [43].

### 2.2.3 Sampling from a Distribution

In summary, a probability distribution specifies the probability of each possible outcome of a random (discrete or continuous) variable. However, often an actual outcome is required, that is, a specific value needs to be chosen at random, with a probability of that value (and all other values) dictated by the given probability distribution. The process of choosing an actual specific value from a probability distribution is called *sampling* or *drawing from the distribution*.

To model an actual sampling process, one employs the *cumulative distribution function (cdf)*.

The cumulative distribution function of a probability function $f$ is defined for each value $x$ in the domain $[u, v]$ of $f$ as the partial sum $\sum_{y=u}^{x} f(y)$ if $f$ is a pmf (for discrete random variables), and $\int_u^x f(y)dy$ if $f$ is a pdf (for continuous variables). The value of $cdf(x)$ for a given $x$ is therefore defined as $P_f(X \leq x)$. Once the *cdf* of a probability function $f$ is known, sampling from $f$ proceeds in two simple steps: (1) Pick uniformly a random value $r$ in the interval $[0, 1]$; (2) Find the smallest $x$ for which it holds $cdf(x) \geq r$.

### 2.2.4 Bayesian Modeling and Generative Models

The basics of probability theory will be extensively used throughout the thesis. Moreover, a large body of the thesis will rely on the knowledge induced from topic models, which are, broadly speaking, a sub-class of probabilistic generative graphical models. Therefore, we provide a brief introduction to this class of models.

**Generative Models.** A *generative model* is a model for randomly generating observable data, typically given some *hidden parameters* or a latent structure. Generative models contrast with discriminative models. While a generative model is a full probabilistic model of all variables, a discriminative model provides a model only for the target variables conditional on the observed variables. Generative models are typically more flexible than discriminative models in expressing dependencies in complex learning tasks. Moreover, unlike generative models, most discriminative models are inherently supervised and cannot be easily extended to unsupervised learning which is the setting in which this thesis is situated.

**Bayesian Networks.** A *graphical model* is a probabilistic model for which a graph denotes the conditional dependency structure between random variables. A Bayesian network is a type of a probabilistic graphical model. It is a compact representation of a probability distribution that usually happens to be too large to be handled using traditional specifications from probability and statistics such as tables and equations. More technically, a Bayesian network is simply a set of random variables for which the conditional dependencies between variables are depicted by a directed acyclic graph. Bayesian networks are established as an omnipresent tool for modeling and reasoning under uncertainty, and provide a standardized inference "toolkit". Many applications are easily reduced to Bayesian network inference. It allows one to capitalize on Bayesian network algorithms instead of inventing specific algorithms for each new application.

For a more detailed overview of graphical models and Bayesian networks with standard examples, we refer the interested reader to the relevant literature [156, 61, 62].

### 2.2.5 Latent Variables and their Estimation

In a Bayesian network, there exist three different types of variables or nodes in the network: *observed*, *initialized* and *latent* (or *hidden*). Observed nodes are nodes for which the current value is known as they are observable. Initialized nodes have a value which is not observed, but chosen to be a specific value. Finally, latent nodes are variables whose state can not be observed; their existence and value can only be inferred by observing the outcome of the observed variables.

Imagine a graphical model, describing a process with observable nodes $X$ and latent nodes $\Theta$ where the dependency probabilities of $X$ on $\Theta$ are unknown. The process runs several times, which yields a collection of observed variables $x$. Is it possible to use these observations to estimate the values $\theta$ of variables $\Theta$? The best guess that can be made about $\theta$ is to claim that, since it generated the observations $x$, this $x$ is the most likely outcome of $\theta$. We are looking for the value of $\theta$ that gives the highest probability for $P(x|\theta)$ (also called $\mathcal{LK}(\theta|x)$ or the *likelihood* of $\theta$ given $x$). Formally, this is the value $\theta^*$ for which $\theta^* = \arg\max_\theta P(x|\theta)$. This estimation technique is called *Maximum Likelihood Estimation* (MLE). If $\theta$ has a known prior distribution (i.e., $P(\theta)$ is known), it is called *Maximum A Posteriori Estimation* (MAPE). We may then write $\theta^* = \arg\max_\theta P(x|\theta)P(\theta)$. To calculate the MLE, several options are possible. If a model is relatively simple, the maximum can be searched analytically, using the equation $\frac{dP(x|\theta)}{d\theta} = 0$. When using MAPE, $\arg\max_\theta P(x|\theta)P(\theta)$ needs to be calculated. For further clarifications and examples of the two estimation techniques, we refer the reader to [125].

An important notion in this calculation is the *conjugate prior*. In short, if the conditional dependency of variable $X$ on variable $\Theta$ follows a distribution of type $PF_1$, and $\Theta$ has a prior distribution of type $PF_2$, then $PF_2$ is the conjugate prior of $PF_1$ if the posterior distribution of $\Theta$ also follows a type $PF_2$ distribution. We may formally define:

$$P(X|\Theta) \sim PF_1 \quad \wedge \quad P(\Theta) \sim PF_2 \quad \wedge \quad P(\Theta|X) \sim PF_2$$

$$\Rightarrow PF_2 \text{ is the conjugate to } PF_1$$

An advantage of using conjugate priors lies in the ease of calculation. Using a non-conjugate prior introduces complex functions, which optimum may not be

obtained analytically. Another reason why conjugate priors are important is the conservation of distribution. Assume that the data is considered to follow a specific distribution. While posterior knowledge may affect the *belief* in the data, it should not change the structure, that is, the distribution type of the data itself. There are several important combinations of distributions and conjugate priors which are well known and exploited in the literature [217]. For instance, for a Bernoulli and binomial distribution, the conjugate prior is the Beta distribution, while for a multinomial distribution, the conjugate prior is the Dirichlet distribution.[1]

Finally, the MLE and MAPE calculations are not always computationally feasible. In complex models, many variables interact with each other, or the number of variables and parameters explodes, making the exact analytical estimation computationally infeasible. Therefore, several alternative estimation schemes were designed. First, there is the *Expectation-Maximization (EM)* algorithm [70]. EM is a deterministic iterative algorithm well suited for dealing with incomplete data, which alternates between: (1) performing an expectation step (E-step), where a function for the expectation of the (log)-likelihood evaluated using the current estimate for the parameters is created (i.e., posterior probabilities are computed for the latent variables based on the current estimates of the parameters), and (2) a maximization step (M-step), in which parameters are computed by maximizing the expected (log)-likelihood found during the previous E-step (i.e., the parameters are re-estimated in order to maximize the likelihood function). Following that, these parameter estimates are then again employed to determine the distribution of the latent variables in the next E step and the process converges to a (local) optimum. Another family of deterministic estimation schemes are *variational Bayesian methods* [140, 90], in which a simplified model is used, constrained to produce results with maximum similarity to the real results. Variational Bayes can be seen as an extension of the EM algorithm from MAPE estimation of the single most probable value of each parameter to fully Bayesian estimation which computes (an approximation to) the entire posterior distribution of the parameters and latent variables [140]. Finally, another alternative estimation method called *Gibbs sampling* [103, 24] will be described in more detail in the upcoming chapters, since it will be heavily used throughout the thesis (due to its intuitive interpretation and the ease of implementation). In short, Gibbs sampling is a randomized estimation model which iteratively samples variables and updates their values until convergence. More on Gibbs sampling follows in chapter 4.

The main insights from probability theory and Bayesian modeling will be heavily used in all parts of the thesis. The basic understanding of probability

[1] Again, for a more elaborate introduction to distributions, conjugate priors and Bayesian estimation, we refer the curious reader to the excellent Heinrich's overview [125].

distributions, sampling, Bayesian networks, the difference between observed, initialized and hidden variables, conjugate priors, etc., is necessary to further understand the essentials of multilingual probabilistic topic modeling discussed in part I. Since the output of multilingual topic models is further utilized in part III and part IV of the thesis (see again fig. 1.1), the basic understanding of these concepts is transitively required to comprehend the thesis as a whole.

## 2.3 Statistical Language Modeling

Another important aspect that needs basic clarifications before being introduced later in the thesis is that of *statistical language modeling*. We extensively employ probabilistic language models in information retrieval in part IV of the thesis.

A statistical language model assigns a probability to a sequence of $N$ words $P(w_1, \ldots, w_N)$ (also called an *N-gram*) by means of a probability distribution. This probability is typically decomposed into its component probabilities using the *chain rule* as follows:

$$P(w_1, \ldots, w_N) = P(w_1) \cdot P(w_2|w_1) \cdot \ldots \cdot P(w_N|w_{N-1}, \ldots, w_1) \tag{2.12}$$

Due to computational reasons, language models typically have a limited memory. It is assumed that the probability of observing the $n$-th word may be estimated by including only the preceding $n-1$ as the context history in the computation of the final probability:

$$P(w_1, \ldots, w_N) = \prod_{i=1}^{N} P(w_i|w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{N} P(w_i|w_{i-(n-1)}, \ldots, w_{i-1}) \tag{2.13}$$

The conditional probabilities are computed from $n$-gram frequency counts in the data:

$$P(w_i|w_{i-(n-1)}, \ldots, w_{i-1}) = \frac{count(w_{i-(n-1)}, \ldots, w_{i-1}, w_i)}{count(w_{i-(n-1)}, \ldots, w_{i-1})} \tag{2.14}$$

These language models are called $n$-gram language models. For instance, if $n = 2$, we talk about a 2-gram (or bigram) language model. The probability $P(w_1, \ldots, w_N)$ is then approximated as:

$$P(w_1, \ldots, w_N) \approx P(w_1) \cdot P(w_2|w_1) \cdot \ldots \cdot P(w_N|w_{N-1}) \tag{2.15}$$

Another, even simpler model is a *unigram* language model. In this model, all conditional dependencies are removed, and the probability of a sequence of words depends only on the probabilities of isolated words:

$$P(w_1, \ldots, w_N) = P(w_1) \cdot P(w_2) \cdot \ldots \cdot P(w_N) \tag{2.16}$$

The unigram model also stands in a strong connection with the *bag-of-words* text representation in which a text is represented as the bag (or the multiset) of its words, disregarding grammar, word order or any syntactic relation. In addition, the unigram language model specifies a multinomial distribution over words.

Statistical language models have found direct application in information retrieval [242, 185], which is investigated further in this thesis. As stated in [185], a common suggestion to users for coming up with good queries is to think of words that would likely appear in a relevant document, and to use those words as the query. The language modeling approach models exactly that idea: a document is relevant to the query if the document is likely to generate the query, and this in turn is likely to happen if the document contains the query words often. In short, the basic language modeling approach builds a probabilistic language model from each document, and ranks documents based on the probability of the model generating the query (more on this in part IV). Language-modeling work in information retrieval typically relies on unigram language models. Retrieval models do not directly depend on the structure of sentences like other tasks such as speech recognition or machine translation, and unigram models are often sufficient to judge the topic of a text [185]. Unigram language models are often *smoothed* to avoid zero-value probabilities $P(w_i)$ for query terms. Smoothing refers to assigning some of the total probability mass to unseen words (or $N$-grams).

Statistical language modeling is a vibrant field of research per se, but its current advances, limitations and more elaborate modeling principles fall way out of the scope of this thesis. The interested reader may check several survey papers and/or handbooks [46, 141, 260, 108, 20, 71]. Besides information retrieval, language models have been used in a wide range of applications such as speech recognition [138], machine translation [37] or spelling correction [145].

## 2.4 Text Preprocessing

Prior to any further operations with text data, the data typically has to be preprocessed using several standardized text analysis techniques. Here, we cover only a subset of these techniques which are used later in the thesis.

**Tokenization.** Typically, it is the first step required for any further text analysis. The task of tokenization refers to the automatic detection of words' boundaries in a text document, that is, it divides the text document into individual *word tokens*. A basic way to achieve tokenization is to rely on clues such as spaces or punctuation. However, this basic approach generates issues with expressions such as named entities or collocations that contain more than

one word. A standard solution to address this problem is to use special lists of named entities or rely on word co-occurrence to decide which expressions should not be split by a tokenizer. Tokenization is a solved problem for many languages (e.g., English, Spanish, Dutch), but it still remains only a partially solved problem for some languages (e.g., Chinese). Another task closely related to tokenization is *sentence boundary detection*, where the aim is to extract individual sentences again using clues such as punctuation of capitalization.

**Stop Words Removal.** Words that occur very frequently and do not bear any semantic information may negatively affect final results of NLP and IR systems [245]. These words are known as *stop words* and are typically filtered out prior to any additional text processing in NLP and IR in a process called *stop words removal*. Typically, a list of such words called *stop words list* is used for filtering. Stop words removal simply filters out all words from a text document that are contained in the stop words list. Example stop words in English include very frequent function words such as *a*, *the*, *is*, *which*, *that*, etc.

**Part-of-Speech Tagging.** Part-of-speech or POS tagging assigns part-of-speech (POS) labels or *tags* to words according to their different grammatical functions or categories (e.g., nouns, verbs, adjectives, adverbs, pronouns). A simple example follows:

*He*/`PRP` *saw*/`VBD` *a*/`DT` *girl*/`NN` *with*/`IN` *a*/`DT` *cat*/`NN` .

In the example above, each word is tagged by a label addressing its grammatical category. `PRP` denotes a personal pronoun, `VBD` a verb in the past tense, `DT` a determiner, `IN` a preposition, and `NN` a singular noun. The sets of POS tags are typically hand-built (e.g., the Penn Treebank Tagset for English [187], also used in the example above) and differ across languages. However, recent trends in POS tagging strive towards an universal language-independent set of POS tags [239] and completely unsupervised language-independent POS systems [63]. POS tagging is a sub-field of NLP research on its own and it is well beyond the scope of this thesis.

**Lemmatization.** Another important preprocessing step involves a shallow morphological analysis of the given text data. If one operates with individual words without any additional morphological analysis, a problem of *data sparsity* may arise due to the fact that some words are actually only different variants (e.g., they differ in tense, person, gender or number) of the same root word (e.g., consider words *build*, *builds*, *building*, *built*, which are all variants of the same root word *build*). In order to address this issue, a common preprocessing step is to perform a morphological analysis of text data. This analysis refers to the process of finding stems and affixes for words, and then mapping them to common roots by stemming and lemmatization. A *stem* is defined as the

major and the smallest meaningful morpheme of the word and the one that carries the word's meaning. Stemming techniques are heuristic-based algorithms that remove typical prefixes and suffixes (for instance, a suffix *-s* in English for third-person singular) and leave only the stem of the word. However, due to their heuristic nature they often remove too much information, and the process results in stems without any meaning at all. In order to tackle this issue, a dictionary-based approach to morphological analysis called *lemmatization* always results in dictionary forms of the words called *lemmas*.

We have applied a tokenizer, a POS tagger and a lemmatizer from the TreeTagger project [267] which may be found online.[2] We have used stop words lists provided for the Snowball project, which may also be acquired online.[3]

## 2.5 Conclusion

In this chapter, a short overview of fundamental tools has been provided, which are the basis for the further modeling and development in this thesis. We have presented a short introduction to probability theory which serves as the main cornerstone for all further modeling and statistical probabilistic representations of text (e.g., words or documents) discussed later in the thesis. We have introduced and tackled key concepts of probability theory such as random variables, conditional probabilities, joint probabilities, probability distributions, prior and posterior distributions, and sampling from a distribution. Following that, a brief introduction to graphical models and Bayesian networks has been provided, necessary to understand the basics of (multilingual) probabilistic topic modeling, a probabilistic modeling principle which serves as the backbone of this thesis. Bayesian networks are simply descriptions of stochastic processes, networks of conditionally dependent variables through which information is propagated to produce a random outcome. Observed outcomes allow for the estimation of the probabilities even for variables that cannot be observed (i.e., they are latent or hidden). The Bayesian framework allows for discovering a latent structure underlying textual data collections. For instance, latent topics may be observed as a hidden knowledge behind the observed text data which is involved in the generation of the actual observed text data.

We have also covered the very basics of statistical language modeling, necessary to fully understand information retrieval models discussed in part IV. Following that, we have provided a short overview of core text preprocessing techniques: tokenization, stop words removal, part-of-speech tagging and lemmatization, the techniques that we utilized in this thesis to prepare our text corpora for further processing.

[2] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
[3] `http://snowball.tartarus.org/`

# Part I

# Multilingual Text Mining

# 3

# Multilinguality and Multilingual Data

*Come, let us go down and confuse their language so they will not be able to understand each other.*
— Genesis 11:7

## 3.1 Introduction: Why Multilingual Data?

Since the Web, and consequently the entire globally connected world, have truly become a multilingual data-driven environment, a need to successfully navigate through that sea, or rather ocean of multilingual information becomes more pressing than ever. As already discussed in chapter 1, we require fully *data-driven cross-lingual* tools. Consequently, to build such data-driven cross-lingual tools, we require a primordial resource - the multilingual data itself. Additionally, a large layer of the deep ocean of multilingual information is still in pure text format. Following this line of thinking, and concerning the fact that this thesis addresses models, representations and applications which operate with text only, in this chapter we provide a short overview of *multilingual text data*. We further motivate why we have decided to work with multilingual data and to build data-driven models in the multilingual setting. We provide a classification and key properties of multilingual text data with example datasets, where the emphasis is on the data that we utilize throughout this work.

In multilingual data, content is typically represented in more than one language, while content in one document is given in only one language. Working with

multilingual data and developing cross-lingual tools provides several advantages over monolingual data and tools:

(a) *Increased functionality* - by making information available to users whose native language is not that of the original information or who cannot easily access such information without an automatic cross-lingual translation assistance or retrieval tool (e.g., the role of bilingual lexicons as discussed in part II and part III, or the role of cross-lingual information retrieval models discussed in part IV).

(b) *More users* - Independent of the actual cross-lingual task, a wider language support implies a wider coverage and a larger number of potential users.

(c) *More available data* - Many datasets come naturally in two or more different languages. Furthermore, focusing only on monolingual aspects of such datasets effectively leads to a loss of important information. Although this is not the focus of our work, it is also worth mentioning that some initial studies [16, 120] show that the exploitation of multilingual information may even lead to increased performance of monolingual models.

(d) *A generalization above monolingual models and tasks* (e.g., cross-lingual semantic similarity models or cross-lingual information retrieval models discussed in part III and part IV are more general than their monolingual variants).

These desirable properties motivate us to investigate cross-lingual models and operate with multilingual data. However, since there is no such thing as free lunch, multilingual data, language-independent text representations and cross-lingual tasks introduce new problems and challenges.

## 3.2 Parallel vs. Comparable Data

Depending on their exact properties, multilingual text data may be divided into two broad categories: (a) *parallel corpora*, and (b) *comparable corpora*. Furthermore, comparable corpora may be divided into finer-grained sub-categories based on the level of *comparability* between texts given in different languages (see later, section 3.2.2).

A critical information when operating with multilingual data concerns the knowledge of *alignment*, that is, the pairing of matched content across languages in a multilingual corpus. The content matching may be observed at different levels of granularity, e.g., the *document alignment* concerns the pairing of related documents, while the *sentence alignment* refers to the pairing of related sentences in a multilingual corpus. In a similar fashion, one may observe phrase

or word alignments. Different algorithms may be used depending on the level of alignment present in a multilingual training corpus. For instance, the algorithms that are tailored to work with sentence-aligned corpora (e.g., the algorithms discussed in part II) are unusable with corpora which provide only document alignments and nothing beyond (e.g., the setting in part III).

### 3.2.1 Parallel Corpora

**Definition 3.1. Parallel corpus.** A *parallel corpus* is a collection of documents given in $l$ different languages, where each document has its exact counterpart, that is, a direct translation in each of the other $l - 1$ languages.

Since matched documents in a parallel corpus are direct translations, and the strict order of sentences is preserved, it is possible to establish a reliable *sentence alignment* between the matched documents. The sentence alignment task is a solved problem in NLP with state-of-the-art models reaching almost perfect performances (see, e.g., [96, 210, 291]), and the majority of popular parallel corpora are deployed with the established sentence alignment links. Alignments of parallel corpora at sentence level are prerequisite for many areas of linguistic research (e.g., statistical machine translation [154]). Furthermore, as all segments in a parallel corpus, without exceptions, have their exact-matching counterparts, a large amount of information can be learned from a parallel corpus. However, providing high-quality exact translations of documents requires a significant human effort and expertise, especially when translations contain a lot of domain-specific jargon words or idiomatic terminology words and expressions. Obtaining such high-quality parallel corpora is an expensive process. Therefore, these corpora are available only for a very restricted set of language pairs and domains. Also, parallel corpora are typically of limited size, and are often not free. In summary, it limits their usage in a general setting.

**Example Corpora.** Perhaps the most famous example of a parallel corpus is the Rosetta Stone. The Rosetta Stone also provides an immediate illustration of the importance of having parallel data as it was the key for deciphering Egyptian hieroglyphs in 1820s. A common source of parallel data are parliamentarian proceedings from an organization such as the European Union, (EU) or countries with more than one official language such as Canada, Hong Kong or Belgium. A historically relevant Hansard corpus with proceedings of the Canadian Parliament in English and French ignited the wave of research in statistical machine translation in the early 1990s. A number of parallel corpora may be found online[1]. We list two example parallel corpora which are used in this work:

[1] e.g., `http://opus.lingfil.uu.se/`

(1) *Europarl* [153]. A collection of documents from the European Parliament proceedings. Due to the nature of proceedings, exact translations must be available in all official communication languages of the EU. The current version 7 of the Europarl corpus comprises documents in 21 languages. In this thesis, we utilize English-Dutch and English-Italian Europarl data. The full statistics of the Europarl corpus are available online.[2] The corpus is extensively used in part II, and it is also used (without the exploitation of the alignments at sentence level) in part III and part IV.

(2) *Moniteur Belge/Belgisch Staatsblad* [307]. A parallel corpus with 5 million (2,4 million unique) French-Dutch aligned sentence pairs in the legislative domain, produced by downloading official documents from the online version of the Belgian Official Journal (Belgisch Staatsblad/Moniteur Belge, an official publication of the Belgian authorities). It covers documents which appeared between 1997 and 2006. The corpus is also freely available online.[3] The corpus is used in appendix A.

### 3.2.2 Comparable Corpora

**Definition 3.2. Comparable corpus.** A *comparable corpus* is a collection of documents with *similar* content which discusses *similar themes* in $l$ different languages, where documents in general are not exact translations of each other.

While parallel corpora are created by one or several persons doing the exact translations, documents from comparable corpora may be created independently and typically originate from a large number of different sources and authors. These documents typically vary in style, length, the usage of vocabulary, the focus on a subject, etc. We detect several degrees of *content comparability* and classify comparable corpora accordingly, similar to Fung and Cheung [92]:

*Type 1.* Every document has a (known) counterpart in each of the other $l-1$ languages, but the counterpart documents are not exact translations of each other.

*Type 2.* Similar to type 1, but in addition the aligned documents only have part of their content in common.

*Type 3.* Documents may or may not have their counterpart documents with similar content in each of the other $l-1$ languages and it is not known which documents possess their counterpart documents and which do not.

[2] `http://www.statmt.org/europarl/`
[3] `http://opus.lingfil.uu.se/MBS.php`

The looser requirements of comparable corpora come at a price, but also with large benefits. Unlike in parallel corpora, exact translations of text segments might not exist in matched documents at all. Moreover, frequencies, sentence orderings and word positions are generally not comparable and may not be used as reliable features. On the bright side, comparable corpora are much cheaper to acquire than parallel corpora, they are available in abundance for many language pairs and domains (e.g., unstructured Web-based user-generated data), they are typically more accessible and more up-to-date. Another advantage of comparable corpora lies in their versatility due to their high availability and a broader scope. Consequently, they have gained a lot of interest from the research community and these corpora slowly begin to find their application in numerous cross-lingual tasks.

**Example Corpora.** For instance, news stories from various sources in the same time frame (type 3) and aligned Wikipedia articles (type 2) often discuss similar events or themes in different languages, but with different focuses. Documents from comparable corpora do not necessarily share all their themes and sub-themes with their counterparts in the other language, but, for instance, Wikipedia articles discussing the same subject, or news stories discussing the same event contain a significant thematic overlap. We could say that such documents in different languages, although inherently non-parallel, are *theme-aligned*. Here, we list an example theme-aligned comparable corpus which is used extensively in this work:

*Aligned Wikipedia articles* [314]. A collection of: (i) 18,898 Italian-English Wikipedia article pairs, (ii) 13,696 Spanish-English Wikipedia article pairs, (iii) 7,612 Dutch-English Wikipedia article pairs. Since the articles are typically written independently and by different authors, rather than being direct translations of each other, there is a considerable amount of divergence between the aligned document pairs. The corpora are collected from Wikipedia *dumps*[4] and are freely available online.[5] The corpora (i)-(iii) are extensively used throughout part III, while the corpus (iii) is also used in part IV.

## 3.3 Conclusions

In this chapter, we have motivated the usage of multilingual text data. We have provided a classification of various types of multilingual text corpora (i.e., parallel vs. comparable) in general and have introduced the multilingual corpora which are used in this thesis in particular. Note that one of our

[4] `http://dumps.wikimedia.org/`
[5] `http://people.cs.kuleuven.be/~ivan.vulic/software/`

research questions (RQ5) motivates the research which aims to answer whether different types of multilingual corpora require different algorithmic principles to obtain optimal results. In this chapter, we have also motivated and displayed the main advantages of comparable corpora (as opposed to parallel corpora). In the next chapter, we will introduce the novel paradigm of multilingual probabilistic topic modeling, which can be utilized to induce knowledge and text representations from such multilingual text resources (related to research question RQ1). Multilingual probabilistic topic models which can operate with non-parallel comparable data will serve as the core of our language pair independent frameworks for modeling cross-lingual semantic similarity (research question RQ2) and cross-lingual information retrieval (research question RQ4) discussed in part III and part IV respectively.

# 4

# Multilingual Probabilistic Topic Modeling

*It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.*

— Arthur Conan Doyle

## 4.1 Introduction

In this chapter, we present a complete systematic and comprehensive overview of the recently developed multilingual probabilistic topic modeling (MuPTM) framework. The overview is the first reported survey on MuPTM in the literature. Due to historical reasons, we start our story from monolingual probabilistic topic models, but later show that these models are actually only special cases of multilingual topic models.

Probabilistic latent topic models such as probabilistic Latent Semantic Analysis (PLSA) [129] and Latent Dirichlet Allocation (LDA) [31] along with their numerous variants are well studied generative graphical models for representing the content of documents in large document collections. They provide a robust and unsupervised framework for performing shallow latent semantic analysis of themes (or topics) discussed in text. The families of these probabilistic latent topic models are all based upon the idea that there exist hidden or latent variables, that is, *topics*, which determine how words in documents have been generated. Fitting such a generative model actually denotes finding

the best set of those latent variables in order to explain the observed data. With respect to that generative process, documents are seen as mixtures of latent topics, while topics are simply probability distributions over vocabulary words. A *mixture* is another name for a discrete multinomial distribution, where the probability values are interpreted as ratios, rather than chances, and representing a document as a realization of mixing topics in a certain proportion is called a *mixture model*. A topic representation of a document constitutes a high-level language-independent view of its content, unhindered by a specific word choice and it improves on text representations that contain synonymous or polysemous words [114].

Probabilistic topic modeling constitutes a very general framework for unsupervised topic mining, and over the years it has been employed in miscellaneous tasks in a wide variety of research domains, e.g., for object recognition in computer vision (e.g., [175, 264, 320]), dialogue segmentation (e.g., [244]), video analysis (e.g., [321]), automatic harmonic analysis in music (e.g., [10, 130]), genetics (e.g., [28]), and others.

Being originally proposed for textual data, probabilistic topic models have also organically found many applications in natural language processing. Discovered distributions of words over topics (further *per-topic word distributions*) and distributions of topics over documents (further *per-document topic distributions*) can be directly employed to detect main themes[1] discussed in texts, and to provide gists or summaries for large text collections (see, e.g., [129, 31, 112, 114]). Per-document topic distributions for each document might be observed as a low-dimensional latent semantic representation of text in a new *topic-document space*, potentially better than the original word-based representation in some applications. In an analogous manner, since the number of topics is usually much lower than the number of documents in a collection, per-topic word distributions also model a sort of dimensionality reduction, as the original word-document space is transferred to a lower-dimensional *word-topic space*. Apart from the straightforward utilization of probabilistic topic models as direct summaries of large document collections, these two sets of probability distributions have been utilized in a myriad of NLP tasks, e.g., for inferring captions for images (e.g., [29]), sentiment analysis (e.g., [196, 293]), analyzing topic trends for different time intervals in scientific literature, social networks and e-mails (e.g., [322, 191, 118]), language modeling in information retrieval (e.g., [323, 330]), document classification (e.g., [31, 158]), word sense disambiguation (e.g., [35]), modeling distributional similarity of terms (e.g., [255, 77]), etc. Lu et al. [182] examine task performance of PLSA and LDA as representative monolingual

[1] To avoid confusion, we talk about *themes* when we address the true content of a document, and *topics* when we address the probability distributions constituting a topic model.

topic models in typical tasks of document clustering, text categorization and ad-hoc information retrieval.

However, all these models have been designed to work with monolingual data, and they have been applied in monolingual contexts only. With the rapid development of Wikipedia and online social networks such as Facebook or Twitter, users have generated a huge volume of multilingual text resources. The user-generated data are often noisy and unstructured, and seldom well-paired across languages. However, as already discussed in sect. 3.2.2, some theme-aligned data (e.g., Wikipedia articles or news data) are abundant in various online sources.

Multilingual probabilistic topic models have recently emerged as a group of unsupervised, language-independent generative machine learning models that can be efficiently utilized on such large-volume non-parallel theme-aligned multilingual data and effectively deal with uncertainty in such data collections. Due to its generic language-independent nature and the power of inference on unseen documents, these models have found many interesting applications. MuPTM aims to model topic discovery from multilingual data in a conceptually sound way, taking into account thematic alignment between documents in document collections given in different languages.

In this chapter, as a representative example, we choose to thoroughly review bilingual LDA, which has been designed as a basic and natural extension of the standard omnipresent LDA model in the multilingual setting, where document-aligned articles in different languages are available (e.g., Wikipedia articles about the same subject in multiple languages). We present a complete and comprehensive overview of that model, all the way up from the conceptual and modeling level, down to its core mathematical foundations, as it could serve as a valuable starting point for other researchers in the field of multilingual probabilistic topic modeling and multilingual text mining. Alternative multilingual probabilistic topic models that are built upon the idea of the standard PLSA and LDA models are presented in a nutshell. These models differ in the specific assumptions they make in their generative processes, as well as in the knowledge that is presupposed before training (e.g., document alignment, prior word matchings or bilingual dictionaries), but all these models share the same conceptual idea and have the ability to discover latent cross-lingual topics from comparable data such as Wikipedia or news. Additionally, all these models output the same basic sets of probability distributions, that is, per-topic word and per-document topic distributions.

Data representation, that is, representations of words and documents by means of the per-topic word distributions and per-document topic distributions (i.e., also an answer to research question RQ1) constitute the core representation

in a variety of applications and models reported in this thesis (part III and part IV). Moreover, the generic nature of such data representations allow for including the topical knowledge in other cross-lingual text mining tasks, which are briefly listed at the end of this chapter. Throughout the thesis, we report the results obtained by BiLDA, but the proposed applications of the MuPTM framework are completely topic model-independent and allow for embedding any other multilingual topic model that outputs the basic set of the per-topic word distributions and per-document topic distributions.

This chapter is structured as follows. We present and define the basic concepts and modeling assumptions related to multilingual probabilistic topic modeling, with a special focus on learning *latent cross-lingual topics* from non-parallel theme-aligned corpora in sect. 4.2. We also provide a generalization of MuPTM and introduce a more general notion of *latent cross-lingual semantic concepts* in sect. 4.3. Following that, in sect. 4.4 the representative bilingual LDA (BiLDA) is presented in its entirety, which includes its generative story, a full explanation of the Gibbs sampling training procedure for the model, the output of the model in terms of per-topic word distributions and per-document topic distributions and the inference procedure. In sect. 4.5, we define and explain several evaluation criteria utilized to compare different probabilistic topic models and list several applications of the MuPTM framework. Alternative multilingual probabilistic topic models are presented in sect. 4.6. Sect. 4.7 summarizes conclusions and future work paths.

## 4.2 A General Framework

### 4.2.1 Definitions and Assumptions

**Definition 4.1. Multilingual theme-aligned corpus**. In the most general definition, a *multilingual theme-aligned corpus* $\mathcal{C}$ of $l = |\mathcal{L}|$ languages, where $\mathcal{L} = \{L_1, L_2, \ldots, L_l\}$ is the set of languages, is a set of corresponding text collections $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l\}$. Each $\mathcal{C}_i = \{d_1^i, d_2^i, \ldots, d_{dn_i}^i\}$ is a collection of documents in language $L_i$ with vocabulary $V^i = \{w_1^i, w_2^i, \ldots, w_{wn_i}^i\}$. Collections $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l\}$ are said to be *theme-aligned* if they discuss at least a portion of similar themes. Here, $dn_i$ denotes the total number of documents in document collection $\mathcal{C}_i$, while $wn_i$ is the total number of words in $V^i$. Moreover, $d_j^i$ denotes the $j$-th document in document collection $\mathcal{C}_i$, and $w_j^i$ denotes the $j$-th word in vocabulary $V^i$ associated with document collection $\mathcal{C}_i$.

**Definition 4.2. Multilingual probabilistic topic model**. A *multilingual probabilistic topic model* of a theme-aligned multilingual corpus $\mathcal{C}$ is a set of

semantically coherent multinomial distributions of words with values $P_i(w_j^i|z_k)$, $i = 1,\ldots,l$, for each vocabulary $V^1,\ldots,V^i,\ldots,V^l$ associated with text collections $\mathcal{C}_1,\ldots,\mathcal{C}_i,\ldots,\mathcal{C}_l \in \mathcal{C}$ given in languages $L_1,\ldots,L_i,\ldots,L_l$. $P_i(w_j^i|z_k)$ is calculated for each $w_j^i \in V^i$. The probabilities $P_i(w_j^i|z_k)$ build *per-topic word distributions* (denoted by $\phi_i$), and they constitute a language-specific representation (e.g., a probability value is assigned only for words from $V^i$) of a language-independent latent cross-lingual concept - topic $z_k \in \mathcal{Z}$. $\mathcal{Z} = \{z_1,\ldots,z_K\}$ represents the set of all $K$ latent cross-lingual topics present in the multilingual corpus. Each document in the multilingual corpus is thus considered a mixture of $K$ latent cross-lingual topics from the set $\mathcal{Z}$. That mixture for some document $d_j^i \in \mathcal{C}_i$ is modeled by the probabilities $P_i(z_k|d_j^i)$ that altogether build *per-document topic distributions* (denoted by $\theta$). In summary, each language-independent latent cross-lingual topic $z_k$ has some probability to be found in a particular document (modeled by per-document topic distributions), and each such topic has a language-specific representation in each language (modeled by language-specific per-topic word distributions).

We can interpret def. 4.2 in the following way: each cross-lingual topic from the set $\mathcal{Z}$ can be observed as a latent language-independent concept present in the multilingual corpus, but each language in the corpus uses only words from its own vocabulary to describe the content of that concept (see fig. 4.1 for an illustrative example). In other words, we could observe each latent cross-lingual topic as a set of discrete distributions over words, one for each language. For instance, having a multilingual collection in English, Italian and Dutch and discovering a topic on *Soccer*, that cross-lingual topic would be represented by words (actually probabilities over words) {*player, goal, scorer, ...* } in English, {*squadra (team), calcio (soccer), allenatore (coach), ...* } in Italian, and {*doelpunt (goal), voetballer (soccer player), elftal (soccer team), ...* } in Dutch. We have $\sum_{w_j^i \in V^i} P_i(w_j^i|z_k) = 1$, for each vocabulary $V^i$ representing language $L_i$, and for each topic $z_k \in \mathcal{Z}$. We say that a latent cross-lingual topic is *semantically coherent* if it assigns high probabilities to words that are semantically related. Def. 4.2 is predominant in the MuPTM literature (e.g., see [67, 203, 241]).

Zhang et al. [333] provide an alternative definition of a multilingual topic model, but we will show that their definition is equivalent to def. 4.2 after a partition over the languages is performed. Namely, the whole multilingual corpus is observed as a mixture of latent cross-lingual topics from $\mathcal{Z}$. They then define a latent cross-lingual topic $z_k \in \mathcal{Z}$ as a semantically coherent multinomial distribution over all the words in all the vocabularies of languages $L_1,\ldots,L_i,\ldots,L_l$, and $P(w_j|z_k)$ gives the probability of any word $w_j \in \{V^1,\ldots,V^i,\ldots,V^l\}$ to be generated by topic $z_k$. In this case, we have

$\sum_{i=1}^{l} \sum_{w_j \in V^i} P(w_j|z_k) = 1$. The language-specific representation for language $L_i$ of topic $z_k$ is then obtained by retaining only probabilities for words which are present in its own vocabulary $V^i$, and normalizing those distributions. For a word $w_j^i \in V^i$, we have $P_i(w_j^i|z_k) = \frac{P(w_j^i|z_k)}{\sum_{w_j \in V^i} P(w_j|z_k)}$. After the partition over languages and normalizations are performed, this definition is effectively equivalent to def. 4.2. However, note that their original definition is more general than def. 4.2, but it is also unbalanced over the languages from $\mathcal{L}$ present in $\mathcal{C}$, that is, words from the languages that are more present in the original corpus $\mathcal{C}$ might dominate the multinomial per-topic word distributions. By performing the partition and normalization over the languages, that imbalance is effectively removed.

**Definition 4.3. Multilingual probabilistic topic modeling**. Given a theme-aligned multilingual corpus $\mathcal{C}$, the goal of *multilingual probabilistic topic modeling* or *latent cross-lingual topic extraction* is to learn and extract a set $\mathcal{Z}$ of $K$ latent language-independent concepts, that is, *latent cross-lingual topics* $\mathcal{Z} = \{z_1, \ldots, z_K\}$ that optimally describe the observed data, that is, the multilingual corpus $\mathcal{C}$. Extracting latent cross-lingual topics actually implies learning *per-document topic distributions* for each document in the corpus, and discovering language-specific representations of these topics given by *per-topic word distributions* in each language (see def. 4.2).

This shared and language-independent set of latent cross-lingual topics $\mathcal{Z}$ serves as the core of unsupervised *cross-lingual text mining* and *cross-lingual knowledge transfer* by means of multilingual probabilistic topic models. It is the cross-lingual connection that bridges the gap across documents in different languages and transfers knowledge across languages in case when translation resources and labeled instances are scarce or missing. The trained multilingual probabilistic topic model may be further inferred on unseen documents.

**Definition 4.4. Inference of a multilingual topic model**. Given an unseen document collection $\mathcal{C}_u$, the inference of a multilingual topic model on the collection $\mathcal{C}_u$ denotes learning topical representations of the unseen documents $d_u \in \mathcal{C}_u$, that is, acquiring per-document topic distributions for the new documents based on the previous output of the model.

**Definition 4.5. Cross-lingual knowledge transfer**. Knowledge transfer in general refers to transferring knowledge learned from one corpus to another corpus, which was unavailable during the learning procedure. *Cross-lingual knowledge transfer* is characterized by the fact that corpora are present in more than one language.

In order to recapitulate the key concepts and intuitions behind multilingual probabilistic topic modeling that have been introduced in this section, we provide

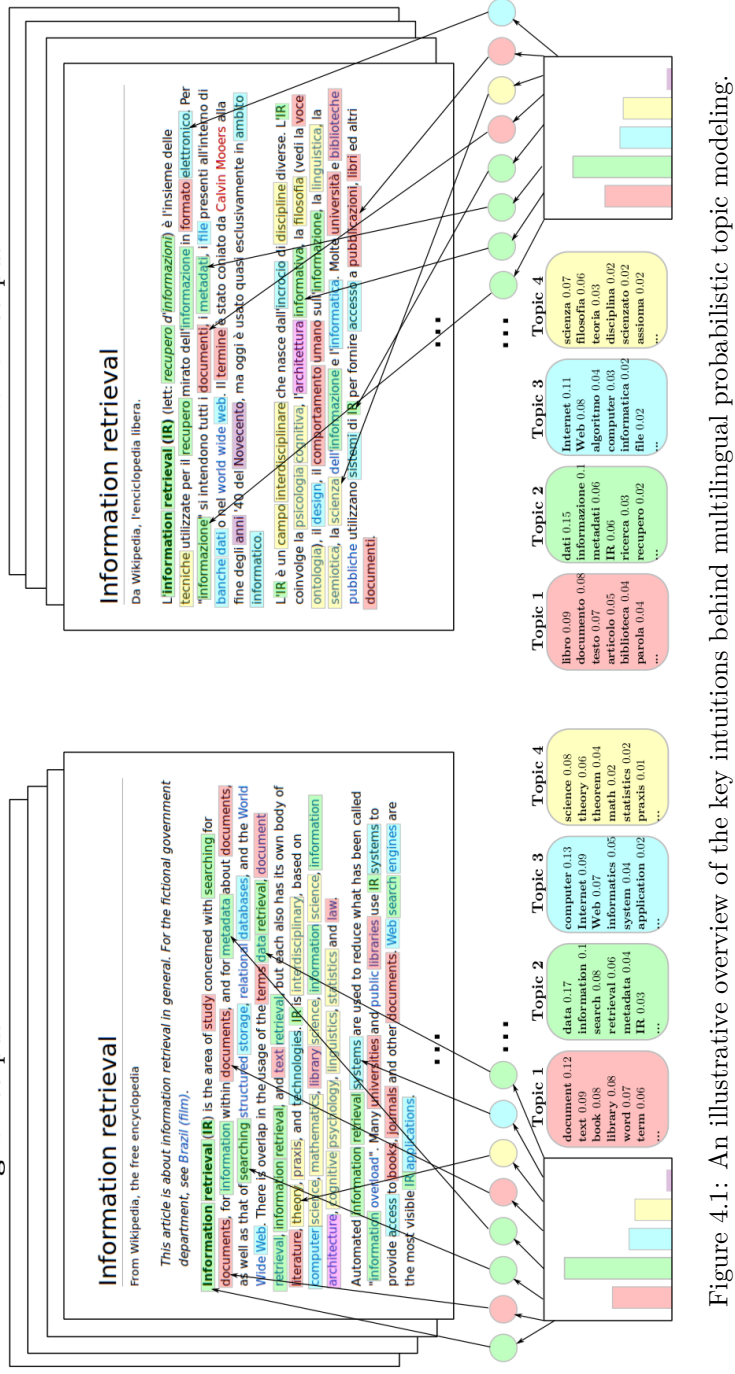



Figure 4.1: An illustrative overview of the key intuitions behind multilingual probabilistic topic modeling.

an illustrative overview in fig. 4.1. Here, each document is represented as a mixture of latent cross-lingual topics (*per-document topic distributions*, presented by histograms), where some latent cross-lingual topics are more important for the particular document. These cross-lingual topics are language-independent concepts, but each language provides a language-specific interface to each cross-lingual topic. In other words, each cross-lingual topic is modeled as a distribution over vocabulary words in each language (*per-topic word distributions*, presented by rounded rectangles). Each document is then generated as follows. First, choose the per-document topic distribution and, according to the distribution, for each word position choose a topic assignment (the colored circles). Following that, according to per-topic word distributions in that language, choose the specific word from the corresponding latent cross-lingual topic that will occur at that word position. Documents that discuss similar themes tend to have similar distributions over cross-lingual topics (the colored bars), but when we operate in the multilingual setting, different per topic-word distributions (the rounded rectangles) are used to generate the observed words in the documents. The generative process does not make any assumptions about syntax, grammar and word order in general, as it assumes that each word is *independently and identically distributed (iid)*, that is, drawn independently from the same distribution (the *bag-of-words assumption)*. Extending the models beyond the bag-of-words assumption (or rather restriction) is possible, but it will not be covered in this work. The figure represents an illustrative toy example, and it is not based on real data.

Additionally, following the assumptions and general definitions provided in this section, *monolingual probabilistic topic models* such as PLSA [129, 128] and LDA [31] could be interpreted as a degenerate special case of multilingual probabilistic topic models where only one language is involved, and all the definitions and assumptions remain the same. Therefore, all models further developed in this thesis that rely on MuPTM (or, more generally, on the notion of *latent cross-lingual concepts*), and which are presented in the multilingual settings are also completely valid and usable in the monolingual settings.

## 4.3 A More General Framework: Latent Cross-Lingual Concepts (Intermezzo)

The latent cross-lingual topics presented in sect. 4.2.1 constitute only one possibility when the aim is to detect and induce a *latent semantic structure* from multilingual data, that is, to extract *latent cross-lingual concepts* that are hidden within the data. Latent cross-lingual concepts may be interpreted as language-independent semantic concepts present in a multilingual corpus

(e.g., document-aligned Wikipedia articles in English and Spanish) that have their language-specific representations in different languages. To repeat, for instance, having a multilingual collection in English, Spanish and Croatian, and discovering a latent semantic concept on *Basketball*, that concept would be represented by words (actually probabilities over words) {*player, ball, coach,...*} in English, {*pelota (ball), jugador (player), partido (match),...*} in Spanish, and {*trener (coach), razigravač (playmaker), doigravanje (playoff),...*} in Croatian.

These $K$ semantic concepts span a latent cross-lingual semantic space. Each word $w$ may be represented in that latent semantic space as a $K$-dimensional vector, where each vector component is a *conditional concept probability score* $P(z_k|w)$. In other words, each word is actually represented as a multinomial probability distribution over the induced latent cross-lingual semantic concepts. Moreover, each document $d$, regardless of its actual language, may be represented as a multinomial probability distribution, a mixture over the same induced latent cross-lingual semantic concepts $P(z_k|d)$.

The description and the work conducted in this thesis rely on the multilingual probabilistic topic modeling framework as discussed in this chapter, but we emphasize that all the work described in this thesis is independent of the actual method used to induce the latent cross-lingual concepts. The reader has to be aware of the fact that the developed models for modeling cross-lingual semantic similarity (part III) and cross-lingual information retrieval (part IV) which are described in the following chapters are generic and model-independent as they allow the usage of all other models that compute probability scores $P(z_k|w)$ and $P(z_k|d)$. Besides MuPTM, a number of other models may be employed to induce the latent cross-lingual concepts. For instance, one could use cross-lingual Latent Semantic Indexing [81], probabilistic Principal Component Analysis [292], or a probabilistic interpretation of non-negative matrix factorization [165, 101, 75] on concatenated documents in aligned document pairs. Other more recent models include matching canonical correlation analysis [117, 64] or other families of multilingual topic models [91].

## 4.4 Bilingual Latent Dirichlet Allocation (BiLDA)

### 4.4.1 An Overview of the Model

Without loss of generality, from now on we deal with *bilingual probabilistic topic modeling*. We work with a *bilingual corpus* and present cross-lingual tasks in the *bilingual setting*. For bilingual corpora we introduce the source language $L_S$ (further with indices $S$) and the target language $L_T$ (further with indices $T$).

We will show that all the definitions and assumptions may be easily generalized to a setting where more than two languages are available.

**Main Assumptions.** Bilingual Latent Dirichlet Allocation (BiLDA) is a bilingual extension of the standard LDA model [31], tailored for modeling parallel document collections, and/or comparable bilingual document collections which are theme-aligned, but loosely equivalent to each other. An example of such a document collection is Wikipedia in two languages with paired articles. BiLDA has been independently designed by several researchers [221, 67, 203, 241]. Unlike LDA, where each document is assumed to possess its own document-specific distribution over topics, the generative process for BiLDA assumes that each *document pair* shares the same distribution of topics. Therefore, the model assumes that we already possess *document alignments* in a corpus, that is, links between paired documents in different languages in a bilingual corpus. This assumption is certainly valid for multilingual Wikipedia data, where document alignment is established via cross-lingual links between articles written in different languages. These links are provided by the nature of the Wikipedia structure. Cross-lingual document alignment for news crawled from the Web is also a well-studied problem. Since the establishing of cross-lingual links between similar documents is not the focus of the research reported in this thesis, these algorithms are not elaborated in the thesis, but we refer the curious reader to the literature (see, e.g., [302, 252, 284, 215, 311]).

**Definition 4.6. Paired bilingual corpus**. A *paired bilingual document corpus* is defined as $\mathcal{C} = \{d_1, d_2, \ldots, d_r\} = \{(d_1^S, d_1^T), (d_2^S, d_2^T), \ldots, (d_r^S, d_r^T)\}$, where $d_j = (d_j^S, d_j^T)$ denotes a pair of linked documents in the source language $L_S$ and the target language $L_T$, respectively.

The goal of bilingual probabilistic topic modeling is to learn for a (paired or non-paired) bilingual corpus a set of $K$ latent cross-lingual topics $\mathcal{Z}$, each of which defines an associated set of words in both $L_S$ and $L_T$.

**The Model.** BiLDA can be observed as a three-level Bayesian network that models document pairs using a latent layer of shared topics. Fig. 4.2 shows the graphical representation of the BiLDA model in plate notation, while alg. 4.1 presents its generative story.

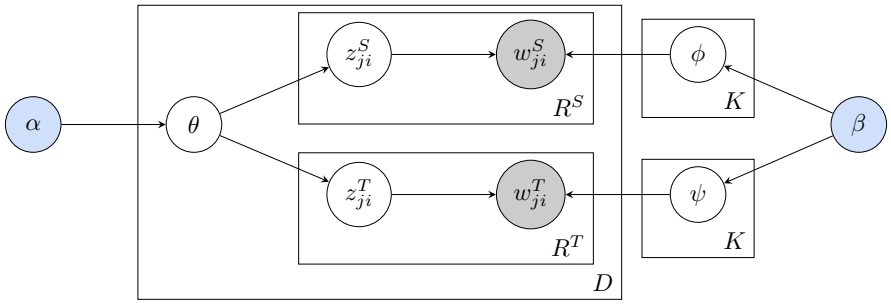



Figure 4.2: Graphical representation of the bilingual LDA (BiLDA) model in plate notation. $R^S$ and $R^T$ denote lengths of the source document and the target document in terms of word tokens for each aligned document pair.

**Algorithm 4.1**: GENERATIVE STORY FOR BILDA

**initialize**: (1) set the number of topics $K$;
(2) set values for Dirichlet priors $\alpha$ and $\beta$;
**sample**: $K$ times $\phi \sim Dirichlet(\beta)$;
**sample**: $K$ times $\psi \sim Dirichlet(\beta)$;
**foreach** *document pair* $d_j = \{d_j^S, d_j^T\}$ **do**
  sample $\theta_j \sim Dirichlet(\alpha)$;
  **foreach** *word position* $i \in d_j^S$ **do**
    sample $z_{ji}^S \sim Multinomial(\theta)$;
    sample $w_{ji}^S \sim Multinomial(\phi, z_{ji}^S)$;
  **end**
  **foreach** *word position* $i \in d_j^T$ **do**
    sample $z_{ji}^T \sim Multinomial(\theta)$;
    sample $w_{ji}^T \sim Multinomial(\psi, z_{ji}^T)$;
  **end**
**end**

BiLDA takes advantage of the assumed topical alignment at the level of linked documents by introducing a single variable $\theta$ (see sect. 4.2.1) shared by both documents. $\theta_j$ denotes the distribution of latent cross-lingual topics over each document pair $d_j$. For each document pair $d_j$, a per-document topic distribution $\theta_j$ is sampled from a conjugate Dirichlet prior with $K$ parameters $\alpha_1, \ldots, \alpha_K$. Note that the correct term here should be *per-pair topic distribution* for BiLDA and *per-tuple topic distribution* in case when more than two languages are involved, but we have decided to retain the original name of the distribution in order to draw a direct comparison with standard monolingual LDA.

Then, with respect to $\theta_j$, a cross-lingual topic $z_{ji}^S$ is sampled. Each word $w_{ji}^S$ at

the position $i$ in the source document of the current document pair $d_j$ is then sampled from a multinomial distribution $\phi$ conditioned on the chosen topic $z^S_{ji}$ at position $i$. Similarly, each word $w^T_{ji}$ of the target language[2] is also sampled following the same procedure, but now using the multinomial distribution $\psi$. Note that words at the same positions in source and target documents in a document pair need not be sampled from the same latent cross-lingual topic (for an overview, see again fig. 4.1). The only constraint imposed by the model is that the overall distributions of topics over documents in a document pair modeled by $\theta_j$ have to be the same. In practice, it does not pose a problem when dealing with theme-aligned comparable data such as Wikipedia articles.

**Hyper-Parameters.** According to [114], each hyper-parameter $\alpha_k$ could be interpreted as a prior observation count for the number of times topic $z_k$ is sampled in a document (or document pair) before having observed any actual words. If one is in possession of a certain prior or external knowledge (e.g., document metadata, main themes of a document collections) about the topic importance and the likelihood of its presence in the data, introducing asymmetric priors gives more preference to a subset of the most important topics, which could in the end lead to a better estimated set of output distributions [202, 136]. However, it is often the case that we do not possess any prior knowledge about themes in a text collection (this will be the case throughout this thesis), and then it is reasonable to assume that all topics are a priori equally likely. Therefore, it is convenient to use a symmetric Dirichlet distribution with a single hyper-parameter $\alpha$ such that $\alpha_1 = \ldots = \alpha_K = \alpha$. Similarly, a symmetric Dirichlet prior is placed on $\phi$ and $\psi$ with a single hyper-parameter $\beta$. $\beta$ may be interpreted as a prior observation count of the number of times words in each language are sampled from a topic before any observations of actual words. Placing these Dirichlet prior distributions on multinomial distributions $\theta$, $\phi$ and $\psi$ results in smoothed per-topic word and per-document topic distributions, where the values for $\alpha$ and $\beta$ determine the degree of smoothing. The influence of these hyper-parameters on the overall quality of learned latent topics is a well-studied problem in monolingual settings [11, 182] and it can be generalized to multilingual settings.

**Extending BiLDA to More Languages.** A natural extension of BiLDA that operates with more than two languages, called *polylingual topic model* (PolyLDA) is presented in [203]. A similar model is proposed in [221, 222]. Instead of document pairs, they deal with *document tuples* (where links between documents in a tuple are given), but the assumptions made by their model remain the same. Fig. 4.3 shows the graphical representation in plate notation of the BiLDA

[2]For the sake of simplicity, both words ($w$-s) and topics ($z$-s) are annotated with a corresponding superscript S or T to denote which language they are used in, although one should keep in mind that the topics are shared language-independent latent concepts.

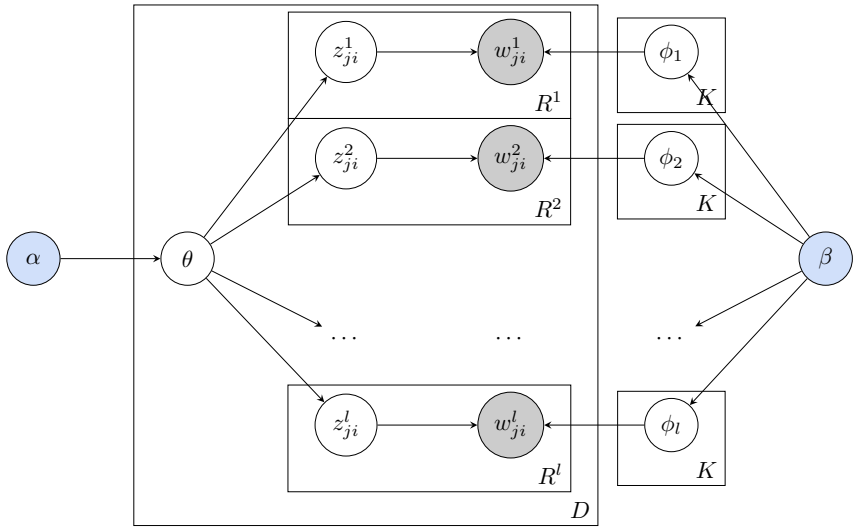



Figure 4.3: Polylingual topic model: The generalization of the BiLDA model which operates with $l$ languages, $l \geq 2$.

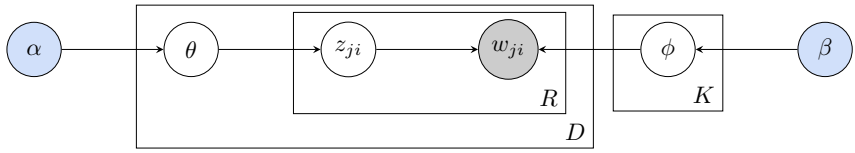



Figure 4.4: Standard monolingual LDA model from Blei et al. [31].

model generalized to $l$ languages, $l \geq 2$, with document tuples $d_j = \{d_j^1, \ldots, d_j^l\}$ and a discrete set of $l$ language-specific per-topic word distributions $\{\phi_1, \ldots, \phi_l\}$ (see sect. 4.2.1).

On the other hand, when operating with only one language, BiLDA or (more generally) PolyLDA is effectively reduced to the standard monolingual LDA model (see fig. 4.4 and compare it with fig. 4.2 or fig. 4.3) [31], that is, the monolingual LDA model is only a degenerate special case of BiLDA and PolyLDA (see sect. 4.2.1).

### 4.4.2 Training: Estimating the BiLDA Model

The goal of training the BiLDA model is to discover the layer of latent cross-lingual topics that describe observed data, i.e., a given bilingual document collection in an optimal way. It means that the most likely values for $\theta$, $\phi$ and $\psi$ have to be found by the training procedure. In simple words, we need to detect and learn which words are important for a particular topic in each language

(that is reflected in per-topic word distributions $\phi$ and $\psi$), and which topics are important for a particular document pair (as reflected in per-document topic distribution $\theta$).

Similarly to the LDA model, the cross-lingual topic discovery for BiLDA is complex and cannot be solved by an analytical learning procedure. There exist a few alternative estimation techniques such as the EM algorithm, variational Bayes estimation or Gibbs sampling (see sect. 2.2.5 in chapter 2). The EM algorithm was used as the estimation method for PLSA [128] and its cross-lingual variant [333]. Variational estimation for the monolingual LDA was used as the estimation technique in the seminal paper by Blei et al. [31]. Other estimation techniques for the monolingual setting include Gibbs sampling [103, 278], and expectation propagation [205, 112].

An extension of the variational method to multilingual settings and its complete formulation for BiLDA were proposed and described by De Smet and Moens [67]. However, due to its prevalent use in topic modeling literature in both monolingual and multilingual setting [33, 203, 135, 312], its ease of implementation and comprehension, as well as its randomized nature (which helps moving away from a local optimum in some cases, unlike with EM and variational Bayes which are deterministic algorithms), we opt for *Gibbs sampling* as the estimation technique for the BiLDA model in all further work and all applications described in this thesis. Therefore, we here provide a detailed overview of the BiLDA training procedure using Gibbs sampling.

**Gibbs Sampling.** Gibbs sampling is a Monte Carlo Markov chain (MCMC) estimation technique. MCMC is a random walk over a Markov chain where each state represents a sample from a specific joint distribution. Starting from a random initial state, the next state is repeatedly sampled randomly from the transition probabilities, and this is repeated until the equilibrium state is reached, in which case states are samples from the joint probability distribution. In Gibbs sampling, it is possible to reach the other states from a given state if only one variable differs in value, while the values of all other variables are held fixed and the transition probabilities are the posterior probabilities for the updated variable. By continuously cycling through each variable until convergence, the Gibbs sampler reaches the equilibrium state. The final samples/estimates are then taken from the full joint distribution [24].

**Algorithm 4.2**: GIBBS SAMPLING: A GENERAL OVERVIEW

**repeat**
  in step $t$;
  choose variable $Y_i$ cyclically from $\mathbb{Y}$;
  sample $Y_i^{t+1} \sim P(Y_i^t | Y_{\neg i}^t)$;
  $Y_{\neg i}^{t+1} = Y_{\neg i}^t$;
**until** *convergence in time t* ;

More formally, alg, 4.2 presents the Gibbs sampling procedure for a Bayesian network $\mathbb{Y}$, where $Y_{\neg i}$ means all $Y$ excluding $Y_i$. $Y_i^t$ is $Y_i$ at time step $t$. For BiLDA in specific, the Gibbs sampling procedure follows these steps presented in alg. 4.3. After the convergence or the equilibrium state is reached, a standard

**Algorithm 4.3**: GIBBS SAMPLING FOR BILDA: AN OVERVIEW

**repeat**
  **consider each word token** in each document in the collection **in turn**;
  **update/estimate the probability** to assign the word token to one of the cross-lingual topics conditioned on all other variables (including all other topic assignments);
  **sample the actual topic assignment** for the word token according to the estimated probabilities;
**until** *convergence/the equilibrium state* ;

practice is to provide estimates of the output distributions as averages over several samples taken in the equilibrium state.

BiLDA requires two sets of formulas to converge to correct distributions:

- One for each topic assignment $z_{ji}^S$ assigned to a word position $i$ that generated word $w_{ji}^S$ in a document given in language $L_S$ in a pair $d_j$.
- One for each topic assignment $z_{ji}^T$ assigned to a word position $i$ that generated word $w_{ji}^T$ in a document given in language $L_T$ in a pair $d_j$.

$\theta$, $\psi$ and $\phi$ are not calculated directly, but estimated afterwards. Therefore, they are integrated out of all the calculations, which actually leaves $z_{ji}^S$-s and $z_{ji}^T$-s as the only latent unknown variables. For the source part (indices $S$) of each document pair $d_j$ and each word position $i$, we calculate the probabilities that $z_{ji}^S$ assumes, as its new values, each of the $K$ possible topic indices $z_k \in \mathcal{Z}$ (indicated by the variable $k$). The actual topic assignment in the sampling step

is then drawn from the current estimate of the probability distribution:

$$\text{sample } z_{ji}^S \sim P(z_{ji}^S = k|\mathbf{z}_{\neg ji}^S, \mathbf{z}_j^T, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta)$$

$$\sim \int_{\theta_j}\int_{\phi} P(z_{ji}^S = k|, \mathbf{z}_{\neg ji}^S, \mathbf{z}_j^T, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta, \theta_j, \phi) d\phi d\theta_j \qquad (4.1)$$

In this formula, $\mathbf{z}_j^T$ refers to all target topic assignments for document pair $d_j$, and $\mathbf{z}_{\neg ji}^S$ denotes all current source topic assignments in $d_j$ excluding $z_{ji}^S$. $\mathbf{w}^S$ denotes all source language word tokens, $\mathbf{w}^T$ all target language word tokens in the entire corpus. Similarly, $\mathbf{w}_{\neg ji}^S$ will denote all word tokens in the corpus excluding $w_{ji}^S$. Sampling for the target side (indices $T$) is performed in an analogous manner:

$$\text{sample } z_{ji}^T \sim P(z_{ji}^T = k|\mathbf{z}_{\neg ji}^T, \mathbf{z}_j^S, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta)$$

$$\sim \int_{\theta_j}\int_{\psi} P(z_{ji}^T = k|, \mathbf{z}_{\neg ji}^T, \mathbf{z}_j^S, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta, \theta_j, \psi) d\psi d\theta_j \qquad (4.2)$$

We further show the derivation of the Gibbs sampler for BiLDA and explain the notation only for the source side of a bilingual corpus and the source language $L_S$ with indices $S$, since the derivation for the target side (with indices $T$) follows in a completely analogous manner.

Starting from eq. (4.1), we can further write:

$$\text{sample } z_{ji}^S \propto \int_{\theta_j}\int_{\phi} P(z_{ji}^S = k|\mathbf{z}_{\neg ji}^S, \mathbf{z}_j^T, \theta, \alpha) \cdot P(w_{ji}^S|z_{ji}^S = k, \mathbf{z}_{\neg ji}^S, \mathbf{w}_{\neg ji}^S, \phi, \beta) d\phi d\theta$$

$$\propto \int_{\theta_j} P(z_{ji}^S = k|\theta_j) \cdot P(\theta_j|\mathbf{z}_{\neg ji}^S, \mathbf{z}_j^T, \alpha) d\theta_j \cdot$$

$$\int_{\phi_k} P(w_{ji}^S|z_{ji}^S = k, \phi_k) \cdot P(\phi_k|\mathbf{z}_{\neg ji}^S, \mathbf{w}_{\neg ji}^S, \beta) d\phi_k$$

Both $\theta$ and $\phi$ have a prior Dirichlet distribution and their posterior distributions are updated with the counter variable $n$ (which counts the number of assigned topics in a document) and the counter variable $v$ (which counts the number of assigned topics in the corpus) respectively (see the explanations of the symbols after the derivation). The expected values ($\int xf(x)dx$) for $\theta$ and $\phi$ become:

$$= E_{Dirichlet(n_{j,k,\neg i}^S + n_{j,k}^T + \alpha)}[\theta_{j,k}] \cdot E_{Dirichlet(v_{k,w_{ji}^S,\neg}^S + \beta)}[\phi_k^{w_{ji}}] \qquad (4.3)$$

Following eq. (4.3), the final updating formulas for both source and target language for the BiLDA Gibbs sampler are as follows:

$$P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta) \propto \frac{n^S_{j,k,\neg i} + n^T_{j,k} + \alpha}{n^S_{j,\cdot,\neg i} + n^T_{j,\cdot} + K\alpha} \cdot \frac{v^S_{k,w^S_{ji},\neg} + \beta}{v^S_{k,\cdot,\neg} + |V^S|\beta} \quad (4.4)$$

$$P(z^T_{ji} = k|\mathbf{z}^T_{\neg ji}, \mathbf{z}^S_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta) \propto \frac{n^T_{j,k,\neg i} + n^S_{j,k} + \alpha}{n^T_{j,\cdot,\neg i} + n^S_{j,\cdot} + K\alpha} \cdot \frac{v^T_{k,w^T_{ji},\neg} + \beta}{v^T_{k,\cdot,\neg} + |V^T|\beta} \quad (4.5)$$

The last two equations use important *counter* variables. The counter $n^S_{j,k}$ denotes the number of times source words in the source document $d^S_j$ of a document pair $d_j$ are assigned to a latent cross-lingual topic $z_k$ (with index $k$), while $n^S_{j,k,\neg i}$ has the same meaning, but not counting the current $w^S_{ji}$ at position $i$ (i.e., it is $n^S_{j,k} - 1$). The same is true for the target side and the $T$ indices.

When a "$\cdot$" occurs in the subscript of a counter variable, this means that the counts range over all values of the variable whose index the "$\cdot$" takes. So, while $n^S_{j,k}$ counts the number of assignments of words $w^S_{ji}$ to one latent topic $z_k$ in $d^S_j$, $n^S_{j,\cdot}$ does so over all $K$ topics in $d^S_j$.

The second counter variable, $v^S_{k,w_{ji},\neg}$ is the number of times a word type whose token appears at position $i$ ($w^S_{ji}$) gets assigned a latent cross-lingual topic $z_k$ in the source side of the entire document collection, but not counting the current $w^S_{ji}$ (i.e., it is $v^S_{k,w_{ji}} - 1$).

Additionally, $\mathbf{z}^S_j$ denotes all latent topic assignments for the source side of the document pair $d_j$, $\mathbf{z}^S_{\neg ji}$ denotes all topic assignments for the source side of $d_j$ but excluding $w^S_{ji}$. $\mathbf{w}^S$ denotes all source words in a corpus, and $|V^S|$ is the number of the source language words in the corpus, that is, the number of words in the vocabulary $V^S$ associated with language $L_S$.

$v^S_{k,\cdot,\neg}$ counts the total number of occurrences of source language words from $V^S$ associated with the topic $z_k$ in the whole corpus, as it is the sum over all possible source language words (a "$\cdot$" appears instead of the $w^S_{ji}$). Again, because of the $\neg$ symbol in the subscript, the current $w^S_{ji}$ is not counted (i.e., the count is then $v^S_{k,\cdot} - 1$).

As can be seen from the first term of eq. (4.4) and eq. (4.5), the document pairs are linked by the count variables $n^S_j$ and $n^T_j$, as both sets of assignments: $z^S_{ji}$ and $z^T_{ji}$ are drawn from the same $\theta_j$. On the other hand, the vocabulary count variables operate only within the language of the word token currently being considered.

### 4.4.3 Output: Per-Document Topic and Per-Topic Word Distributions

With eq. (4.4) and eq. (4.5). each assignment $z_{ji}^S$ and $z_{ji}^T$ of each document pair is sampled and cyclically updated. After a random initialization, usually using a uniform distribution, the sampled values will converge to samples taken from the real joint distribution of $\theta$, $\phi$ and $\psi$, after a time called the *burn-in period*. From a set of complete *burned-in Gibbs samples* of the whole document collection, the final output probability distributions, that is, per-topic word distributions and per-document topic distributions are estimated as averages over these samples.

Language-independent per-document topic distributions provide distributions of latent cross-lingual topics for each document in a collection. They reveal how important each topic is for a particular document. We need to establish the exact formula for per-document topic distributions for documents in an aligned document pair using eq. (4.4) and eq. (4.5):

$$P(z_k|d_j) = \theta_{j,k} = \frac{n_{j,k}^S + n_{j,k}^T + \alpha}{\sum_{l=1}^K n_{j,l}^S + \sum_{l=1}^K n_{j,l}^T + K\alpha} \tag{4.6}$$

Language-specific per-topic word distributions measure the importance of each word in each language for a particular cross-lingual topic $z_k$. Given a source language with vocabulary $V^S$, and a target language with vocabulary $V^T$, and following eq. (4.4), a probability that some word $w_i^S \in V^S$ will be generated by the cross-lingual topic $z_k$ is given by:

$$P(w_i^S|z_k) = \phi_{k,i} = \frac{v_{k,w_i^S}^S + \beta}{\sum_{l=1}^{|V^S|} v_{k,w_l^S}^S + |V^S|\beta} \tag{4.7}$$

The same formula, but now derived from eq. (4.5) is used for the per-topic word distributions ($\psi$) for the target language that specify the probability that some $w_i^T \in V^T$ will be generated by $z_k$:

$$P(w_i^T|z_k) = \psi_{k,i} = \frac{v_{k,w_i^T}^T + \beta}{\sum_{l=1}^{|V^T|} v_{k,w_l^T}^T + |V^T|\beta} \tag{4.8}$$

### 4.4.4 Inference or "What with New Documents?"

Since the model possesses a fully generative semantics, it is possible to train the model on one multilingual corpus (e.g., multilingual Wikipedia) and then infer

it on some other, previously unseen corpus. Inferring a model on a new corpus means calculating per-document topic distributions for all the unseen documents in the unseen corpus based on the output of the trained model. Inference on the unseen documents is performed only one language at a time, e.g., if we train on English-Dutch Wikipedia, we can use the trained BiLDA model to learn document representations, that is, per-document topic distributions for Dutch news stories, and then separately for English news stories.

In short, we again randomly sample and then iteratively update topic assignments for each word position in an unseen document, but now start from the fixed $v$ counters learned in training, and then cyclically update the probability distributions from which the topic assignments are sampled. Since the inference is performed monolingually, dependencies on the topic assignments from another language are removed from the updating formulas. Hence, similar to eq. (4.4), the updating formula for the source language $L_S$ is:

$$P(z^S_{ji} = k|\mathbf{z}^S_{\neg ji}, \mathbf{w}^S, \alpha, \beta) \propto \frac{n^S_{j,k,\neg i} + \alpha}{n^S_{j,\cdot,\neg i} + K\alpha} \cdot \frac{v^S_{k,w^S_{ji}} + \beta}{v^S_{k,\cdot} + |V^S|\beta} \tag{4.9}$$

Learning a multilingual topic model on one multilingual corpus and then inferring that model on previously unseen data constitutes the key concept of *(cross-lingual) knowledge transfer* by means of multilingual probabilistic topic models and that property is extensively utilized in part IV of this thesis as well as in other cross-lingual applications of the model (see sect. 4.7).

## 4.5 Evaluation of Multilingual Topic Models

A simple way of looking at the output quality of a topic model is by simply inspecting top words associated with a particular topic learned during training. We say that a latent topic is *semantically coherent* if it assigns high probability scores to words that are semantically related [106, 220, 204, 277, 5, 72]. It is much easier for humans to judge semantic coherence of cross-lingual topics and their alignment across languages when observing the actual words constituting a topic. These words provide a shallow qualitative representation of the latent topic space, and could be seen as direct and comprehensive word-based summaries of a large document collection. In other words, humans can get the first clue "what all this text is about in the first place".

Besides this shallow qualitative analysis relying on the top words, there are other, theoretically well-founded evaluation metrics for *quantitative* analysis and comparison of different models. In the literature, latent topic models are often evaluated by their perplexity, where the perplexity or "confusion" of a

| FR-EN Topic 17 | NL-EN Topic 55 | IT-EN Topic 73 | ES-EN Topic 52 |
|---|---|---|---|
| moteur | gebouw | rete | dinero |
| *(engine)* | *(building)* | *(network)* | *(money)* |
| voiture | eeuw | chiave | mercado |
| *(vehicle)* | *(century)* | *(key)* | *(market)* |
| automobile | meter | protocollo | precio |
| *(car)* | *(meter)* | *(protocl)* | *(price)* |
| vitesse | kasteel | server | bienes |
| *(speed)* | *(castle)* | *(server)* | *(goods)* |
| constructeur | bisschop | messaggio | valor |
| *(constructor)* | *(bishop)* | *(message)* | *(value)* |
| roue | stad | connessione | cantidad |
| *(wheel)* | *(city)* | *(connection)* | *(amount)* |
| vapeur | gebouwd | client | oferta |
| *(steam)* | *(built)* | *(client)* | *(offer)* |
| puissance | theater | servizion | pago |
| *(power)* | *(theater)* | *(service)* | *(payment)* |
| diesel | museum | indirizzo | impuesto |
| *(diesel)* | *(museum)* | *(address)* | *(tax)* |
| cylindre | tuin | sicurezza | empresa |
| *(cylinder)* | *(garden)* | *(security)* | *(company)* |
| engine | building | link | economic |
| car | court | network | price |
| vehicle | built | display | money |
| fuel | garden | calendar | market |
| speed | museum | client | capital |
| power | palace | key | tax |
| production | construction | server | goods |
| design | theater | protocol | interest |
| diesel | tower | address | demand |
| drive | castle | packet | inflation |

Table 4.1: Randomly selected examples of latent cross-lingual topics represented by top 10 words based on their counts after Gibbs sampling. Topics are obtained by BiLDA trained on Wikipedia for various language pairs: French-English (FR-EN), Dutch-English (NL-EN), Italian-English (IT-EN), and Spanish-English (ES-EN). For non-English words we have provided corresponding English translations. $K = 100$ for all models.

model is a measure of its ability to explain a collection $\mathcal{C}_u$ of unseen documents. The perplexity of a probabilistic topic model is expressed as follows:

$$perp(\mathcal{C}_u) = exp\left(-\frac{\sum_{d\in\mathcal{C}_u}\log\left(\prod_{w\in d}P(w)\right)}{\sum_{d\in\mathcal{C}_u}N^d}\right) \tag{4.10}$$

where $N^d$ is defined as the number of words in a document $d$, $P(w)$ is word $w$'s marginal probability according to a specific model, calculated as $\sum_k P(w|z_k, \Upsilon)$, where $k$ ranges over all $K$ topics in the model, and $\Upsilon$ is the set of the corpus independent parameters of the model. For BiLDA, the parameter set is $\Upsilon = \{\alpha, \beta, \phi, \psi, K\}$. A lower perplexity score means less confusion of the model in

explaining the unseen data, and, theoretically, a better model. A good model with a low perplexity score should be well adapted to new documents and yield a good representation of those previously unseen documents. Since the perplexity measure defines the quality of a topic model independently of any application, it is considered an *intrinsic* or *in vitro* evaluation metric.

Another intrinsic evaluation metric for multilingual probabilistic topic models, named *cross-collection likelihood*, was proposed recently in [333], but that measure also presupposes an existing bilingual dictionary as a critical resource. Additionally, a number of intrinsic quantitative evaluation methods (but for the monolingual settings) are proposed in [318]. Other studies for the monolingual setting focused more on automatic evaluation of semantic coherence (e.g., [44, 220, 204]). However, perplexity still remains the dominant quantitative in vitro evaluation method that is predominantly found in the literature.

Finally, the best way to evaluate multilingual probabilistic topic models is to test how well they perform in practice for different real-life tasks (e.g., document classification, information retrieval), that is, to carry out an *extrinsic ex vivo* evaluation. We later investigate whether there exists a mismatch between the intrinsic and extrinsic evaluation in information retrieval (see sect. 11.5.8). To conclude, we could say that the work reported in this thesis is the first conducted work which evaluates the ability of multilingual topic models to build models of cross-lingual semantic similarity (research question RQ2) and information retrieval (RQ4).

## 4.6 A Short Overview of Other Multilingual Probabilistic Topic Models

Similarly to LDA in the monolingual setting (for which we have already shown that it is only a special case of BiLDA operating with only one language), we believe that bilingual LDA can be considered the basic building block of this general framework of multilingual probabilistic topic modeling. It serves as a firm baseline for future advances in multilingual probabilistic topic modeling. Although MuPTM is a quite novel concept, several other models have emerged over the last years. All current state-of-the-art multilingual probabilistic topic models build upon the idea of standard monolingual PLSA and LDA and closely resemble the described BiLDA model, but they differ in the assumptions they make in their generative processes, and in knowledge that is presupposed before training (e.g., document alignments, prior word matchings or bilingual dictionaries). However, *they all share the same concepts defined in sect. 4.2.1,*

*that is, the sets of output distributions and the set of latent cross-lingual topics that has to be discovered in a multilingual text collection.*

The early approaches (see, e.g., [81, 41]) tried to mine topical structure from multilingual texts using an algebraic model, that is, Latent Semantic Analysis (LSA) and then use the discovered latent topical structure in cross-lingual information retrieval. Artificial "cross-lingual" documents were formed by concatenating aligned parallel documents in two different languages, and then LSA on a word-by-document matrix of these newly built documents was used to learn the lower dimensional document representation. Documents across languages are then compared in that lower-dimensional space.

Another line of work [335, 336] focused on building topic models suitable for word alignment and statistical machine translation operations. Again inspired by monolingual LDA, they have designed several variants of topic models that operate on parallel corpora aligned at sentence level. The topical structure at the level of aligned sentences or word pairs is used to re-estimate word translation probabilities and force alignments of words and phrases generated by the same topic.

However, the growth of the global network and increasing amounts of comparable theme-aligned texts have formed a need for constructing more generic models that are applicable to such large-volume, but less-structured text collections. Standard monolingual probabilistic topic models coming from the families of PLSA and LDA cannot capture and accurately represent the structure of such theme-aligned multilingual text data in a form of joint latent cross-lingual topics. That inability comes from the fact that topic models rely on word co-occurrence information to group similar words into a single topic. In case of multilingual corpora (e.g., Wikipedia articles in English and Dutch) two related words in different languages will seldom co-occur in a monolingual text, and therefore these models are unable to group such pairs of words into a single coherent topic (for examples see, e.g., [33, 135]). In order to anticipate that issue, there have been some efforts that trained monolingual probabilistic topic models on concatenated document pairs in two languages (e.g., [81, 179, 41, 47, 328, 49, 261]), but such approaches also fail to build a shared latent cross-lingual topical space where the boundary between the topic representations with words in two languages is firmly established. In other words, when training on concatenated English and Spanish Wikipedia articles, the learned topics contain both English and Spanish words. However, we would like to learn latent cross-lingual topics for which their representation in English is completely language-specific and differs from their representation in Spanish.

Recently, several novel models have been proposed that remove such deficiency. These models are trained on the individual documents in different languages and

their output are joint latent cross-lingual topics in an aligned latent cross-lingual topical space. The utility of such new topic representations is clearly displayed in part III and part IV. The BiLDA model [67] and its extensions to more than two languages (PolyLDA, [203, 221]) constitute the current state-of-the-art in multilingual probabilistic topic modeling and have been validated in various cross-lingual tasks (e.g., [68, 222, 313]). These models require alignments at document level *a priori* before training, which is easily obtained for Wikipedia or news articles. These document alignments provide hard links between topic-aligned semantically similar documents across languages.

Recently, there has been a growing interest in multilingual topic modeling from unaligned text, again inspired by monolingual LDA. The MuTo model [33] operates with *matchings* instead of words, where matchings consist of pairs of words that link words from the source vocabulary to words from the target vocabulary. These matchings are induced by the matching canonical correlation analysis (MCCA) [117, 64] which ties together words with similar meanings across languages, where similarity is based on different features. Matchings are induced based on pointwise mutual information (PMI) from parallel texts, machine-readable dictionaries and orthographic features captured by, for instance, edit distance. A stochastic expectation-maximization (EM) algorithm is used for training, and their evaluation has been performed on a parallel corpus. A similar idea of using matchings has been investigated in [135]. In their JointLDA model, they also observe each cross-lingual topic as a mixture over these matchings (or *word concepts*, as they name them), where the matchings are acquired directly from a machine-readable bilingual dictionary. JointLDA uses Gibbs sampling for training and it is trained on Wikipedia data. Although these two models claim that they have removed the need for document alignment and are fit to mine latent cross-lingual topics from unaligned multilingual text data, they have introduced bilingual dictionaries as a new critical resource. These machine-readable dictionaries have to be compiled from parallel data or hand-crafted, which is typically more expensive and time-consuming than obtaining alignments for Wikipedia or news data.

Another work that aims to extract latent cross-lingual topics from unaligned datasets is presented in [333]. Their Probabilistic Cross-lingual Latent Semantic Analysis (PCLSA) extends the standard PLSA model [129] by regularizing its likelihood function with soft constraints defined by an external machine-readable bilingual dictionary. They use the generalized expectation maximization (GEM) algorithm [195] for training. Similar to MuTo and JointLDA, a bilingual dictionary is presupposed before training and it is a critical resource for PCLSA. The dictionary-based constraints are the key to bridge the gap between languages by pushing related words in different vocabularies to occur in the same cross-lingual topics. The same relationship between PLSA and LDA [105] in the

monolingual setting is also reflected between their multilingual extensions, PCLSA and BiLDA.

## 4.7 Conclusions and Future Work

In this chapter, we have conducted the first systematic and thorough overview of the current advances in multilingual probabilistic topic modeling. The MuPTM framework will be abundantly utilized in other parts of this thesis. We have provided a precise definition of multilingual probabilistic topic modeling, and have described how to obtain the set of latent cross-lingual topics (or more generally - latent cross-lingual concepts, see sect. 4.3). Multilingual probabilistic topic models which induce these latent cross-lingual topics/concepts in general comprise two basic sets of probability distributions: (1) per-document topic distributions that define topic importance in a document, and (2) per-topic word distributions that define importance of vocabulary words in each language for each cross-lingual topic.

A large part of this thesis then introduces the first applications of this recently developed framework of multilingual probabilistic topic modeling in two major cross-lingual tasks: (1) cross-lingual semantic similarity and bilingual lexicon extraction (part III), and (2) cross-lingual information retrieval (part IV). Since the BiLDA model has been used in these tasks and taking into account that the BiLDA model may be considered as a basic building block in the MuPTM framework (in an analogous manner as LDA in the monolingual setting), in this chapter we have presented a full overview of that multilingual topic model. However, the models presented in the aforementioned parts of this thesis are completely generic, model-independent, language-independent, and language pair independent as long as the supporting topic models output the two basic sets of distributions.

Besides the MuPTM-based frameworks for cross-lingual semantic similarity, bilingual lexicon extraction and cross-lingual information retrieval which were pioneered in this thesis and later sparked even more research interest in applying multilingual topic models in these specific tasks (e.g., [98, 180, 183]), multilingual topic models have found their application in plenty of other cross-lingual NLP and IR tasks. For instance, they have been used in event-based cross-lingual news clustering [67], cross-lingual document classification [222, 68], transliteration mining [253], building comparable corpora [241, 338], cross-lingual keyword recommendation [282], cross-lingual entity linking [334], etc. One line of future work might lead to investigating more applications of the MuPTM framework.

In order to improve the quality of the lower-dimensional topical representations of documents in the multilingual domain, there is a huge number of paths that could be followed. In the same manner as for the natural "LDA to BiLDA" extension, other more sophisticated and application-oriented probabilistic topic models developed for the monolingual setting could be ported into the multilingual setting (e.g., [317, 30, 109]). These extensions include, among others, the use of sentence information or word ordering (using Hidden Markov Models) to yield more coherent topic distributions over documents (e.g., [113, 34]). The use of hierarchical topics (general super-topics connected with more focused sub-topics, see, e.g., [26, 201]) is another interesting field of research in the multilingual setting. Moreover, there is a need to develop multilingual probabilistic topic models that fit data which is less comparable and more divergent and unstructured than Wikipedia or news stories, where only a subset of latent cross-lingual topics overlaps across documents written in different languages. Additionally, the more data-driven topic models should be able to learn the optimal number of topics dynamically according to the properties of training data itself (the so-called *non-parametrized models* [176, 331]), and clearly distinguish between shared and non-shared topics in a multilingual corpus. The emphasis of this thesis however is not on the design and construction of new multilingual topic models, but rather on investigating the utility of the MuPTM framework in the language-independent and language pair independent models of semantic similarity and information retrieval across languages.

Finally, another possibility for future work lies in the expansion of the MuPTM framework from the purely text-based multilingual setting to the *multimodal* setting. Since there exists a significant semantic gap between the "visual words" and textual words, the comparability in the multimodal setting is inherent, and the multilingual probabilistic topic models that operate with comparable text data (see chapter 3) should be transferred to the multimodal setting. They should be capable of dealing with the inherent comparable nature of any dataset consisting of text and some other modality (e.g., video). However, some initial studies [66] have revealed that the multimodal setting serves as a more complex setting, and the additional expansions of the multimodal topic models are needed to effectively handle the existing gap between different modalities.

## 4.8 Related Publications

[1] **I. Vulić**, W. De Smet, J. Tang, and M.-F. Moens. "Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications," in *NIPS 2012 Workshop on Cross-Lingual Technologies*

*(xLiTe)*, Lake Tahoe, Nevada, USA, 7-8 December 2012, 11 pages, NIPS, 2012.

[2] M.-F. Moens and **I. Vulić**. "Monolingual and cross-lingual probabilistic topic models and their application in information retrieval," in *Proceedings of the 35th European Conference on Information Retrieval (ECIR)*, vol. 7814 of *Lecture Notes in Computer Science*, Moscow, Russian Federation, 24-27 March 2013, pp. 875-878, Springer, 2013.

[3] **I. Vulić**, W. De Smet, J. Tang, and M.-F. Moens. "Probabilistic topic modeling in multilingal settings: An overview of its methodology with applications," accepted with minor revisions in *Information Processing & Management*, 2014.

# Part II

# Finding Term Translations in Parallel Data

# Outline of Part II

In part II, the thesis focuses on the multilingual setting, where the existence of parallel corpora is assumed. The goal of this part is to investigate the knowledge induction from parallel corpora, where the focus is on algorithms for bilingual lexicon extraction (BLE), that is, finding term translations from parallel data. Therefore, the research reported here is relevant to research questions RQ2 and RQ5. Since the proposed framework is precision-oriented, this research is also relevant to research question RQ3. The models for BLE from parallel data are well established and typically taken for granted and used off-the-shelf in many NLP and IR applications. However, in this part we show that further improvements in the quality of extracted term translations may be achieved. The reported research in part II is divided into two pieces:

I. As a major contribution of part II, we introduce a new modeling paradigm which relies on the new concept of sub-corpora sampling and present a new language-pair agnostic algorithm called SampLEX for finding term translations in parallel data relying on the paradigm of sub-corpora sampling. The proposed algorithm outscores all other state-of-the-art models for BLE from parallel data.

II. Since the work reported in this part has been conducted within the *TermWise* project, we also present a case study, the final deliverable of the TermWise project in appendix A. There we briefly demonstrate how to put into practice our new approach to finding term translations from parallel data. Namely, our SampLEX algorithm has been integrated as one module in a CAT (computer-assisted translation) tool tailored for assisting translators in the Belgian legislative domain dealing with translations between Dutch and French, two dominant official languages of Belgium.

# 5

# SampLEX: A New Algorithm for Bilingual Lexicon Extraction from Parallel Data

*Less is more only when more is too much.*
— Frank Lloyd Wright

## 5.1 Introduction

*Bilingual lexicon extraction* is the process of automatically acquiring translations of words, phrases or other textual items (in general - *terms*) on the basis of multilingual text resources such as parallel or comparable corpora. Bilingual lexicons serve as an invaluable and indispensable source of knowledge for both end users (as an aid for translators or other language specialists) and many natural language processing and information retrieval systems. It is necessary to build such lexicons manually by hand or extract them automatically from multilingual data. Compiling such lexicons manually is often a labor-intensive and time-consuming task. Due to a scarceness of parallel data for many language pairs and domains, bilingual lexicon extraction from comparable corpora has also gained much interest from the research community recently (see part III of this thesis). However, the "real-world" NLP and IR systems still almost exclusively rely on the knowledge from bilingual lexicons extracted from parallel texts. These lexicons are usually acquired from word translation probabilities of the IBM alignment models [37, 229] or obtained by associative methods such as the log-likelihood ratio (LLR) [82] or the Dice coefficient [74]. They are

then used in systems for extracting parallel sentences from non-parallel corpora [92, 214], bilingual sentence alignment [210], estimating phrase translation probabilities [308], extracting parallel sub-sentential fragments from non-parallel corpora [215], word-level confidence estimation [301], sub-sentential alignment for terminology extraction [171], cross-lingual text classification and plagiarism detection [240], and others. For a more detailed overview of the usage and additional applications of the automatically extracted bilingual lexicons, see later sect. 6.1 in chapter 6.

High accuracy of automatically constructed bilingual word lexicons is the top priority for these systems. Church and Mercer [48] advocate a simple solution of collecting more data in order to utilize statistical and stochastic methods in a more effective way. However, these systems are typically faced with only limited parallel data for many language pairs and domains [252].

In order to tackle these issues, we propose a novel approach built upon the idea of *data reduction* instead of *data augmentation*. The method is directed towards extraction of only *highly reliable translation pairs* from *parallel data of limited size*. It is based on the idea of *sub-corpora sampling* from the original corpus. For instance, given an initial corpus $\mathcal{C}$ of 4 data items $\{I_1, I_2, I_3, I_4\}$, the construction of, say, a sub-corpus $\mathcal{SC} = \{I_2, I_4\}$ may be observed as: (1) sampling items $I_2, I_4 \in \mathcal{C}$ for $\mathcal{SC}$ (hence the term sub-corpora sampling) or (2) removing data items $I_1, I_3$ from the original corpus $\mathcal{C}$, so that $\mathcal{SC} = \mathcal{C} - \{I_1, I_3\}$ (hence the term data reduction). By reducing the size of the initial corpus, we typically decrease frequencies of the words in a newly formed sub-corpus. This simplifies the establishment of potential translation pairs, since that is now reduced to a problem of establishing reliable translational equivalence between low-frequency words. We explain the method for establishing translational equivalence based on the absolute frequency distributions of words in a sub-corpus. We exploit it in the construction of the algorithm for BLE. Moreover, each word exhibits a different distribution over items in each newly built sub-corpus, and it is different from the fixed distribution in the original corpus. It allows us to identify different potential translation pairs in different sub-corpora and then form word translation tables by combining these evidences acquired from different sub-corpora. The key strength of the proposed algorithm is that it takes the entire initial corpus into account, regardless of its size, and at the same time it also benefits from the sampling of a vast number of different subsets/sub-corpora sampled from that initial corpus, and the evidences of potential word translation pairs coming from these sub-corpora.

In the remainder of this chapter, we show that: (1) Bilingual lexicon extraction benefits from the concept of data reduction and sub-corpora sampling - the key intuitions, assumptions and the construction of the SampLEX algorithm are provided in sect. 5.2; (2) The proposed algorithm for BLE removes a lot

of noise from bilingual word lexicons by harvesting only the most accurate translation candidates, and it outperforms other state-of-the-art models for BLE from parallel data which are presented in sect. 5.3. Experimental setup is presented in sect. 5.4. The results on the BLE task are presented in sect. 5.5. Finally, sect. 5.6 lists conclusions and possible paths of future work.

## 5.2 Learning Translation Pairs Using Sub-Corpora Sampling

Sect. 5.1 has already provided a general intuition behind our method for mining translation candidates from aligned corpora. Now, we provide an in-depth description and analysis of our algorithm for bilingual word lexicon extraction. First, we explain the key reasoning that led us to our approach that relies on *data sampling*. Second, we provide the criteria for extracting translation candidates that purely rely on their distributional features, but do not employ any similarity-based measure or hypothesis testing for word association, and finally, we present our algorithm for BLE that processes words of all frequencies in an uniform way.

### 5.2.1 Why Sampling Sub-Corpora?

The foundation of this work is built upon the so-called *Zipfian phenomenon* or the Zipf's law which states that, regardless of the size of a corpus, most of the distinct words occur only a small number of times [339, 340]. For instance, Moore [212] measures that in the first 500,000 English sentences taken from the Canadian Hansards corpus [199], one finds 52,921 distinct words, of which 60.5% occur five or fewer times, and, moreover, 32.8% occur only once. A general solution to mitigate the problem of low-frequency words is by augmenting the amount of input training data. However, that approach leads to a *chicken and egg problem* - adding more data will increase frequencies of the words already present in the corpus, and, accordingly, solve the issue of the low-frequency words, but at the same time, it will introduce many extra words, where some of them were previously out-of-vocabulary. Most of these new words will now be low-frequency words - again we observe the very same Zipfian phenomenon, and the problem of low-frequency words is still present.

Therefore, we have decided to take an opposite path, where "removing" data from the initial corpus (that actually means sampling a sub-corpus with less data items from the original large corpus) and properties of low-frequency words [212, 243] should actually help us detect correct cross-lingual word associations.

By reducing the corpus size, we also decrease frequencies of the words in the corpus. In an extreme case, when the reduced corpus consists of only one sentence, almost all words in that "corpus" will occur only once or twice. Intuitively, for words with higher frequencies, one needs to remove more data, that is, sample a sub-corpus of smaller size, to bring the words down to only a few occurrences in the sub-corpus. We will show that it is easier to establish translational equivalence for low-frequency words.

### 5.2.2 Criteria for Extraction of Translation Pairs

*Translational equivalence* is the similarity in meaning (i.e., the information a word or a term conveys) between a word (or a term) in one language and its translation in another [197]. If a word $w_2^T \in V^T$ exhibits the highest translational equivalence with $w_1^S \in V^S$, it is called a *translation candidate* for $w_1^S$, and the pair $(w_1^S, w_2^T)$ is a *translation pair*. A *potential translation pair* is a pair of words which exhibits a relation of a translation pair in at least one sub-corpus of the original corpus.

Assume that we are in possession of a multilingual corpus $\mathcal{C}$ of $\Omega$ aligned item pairs $\mathcal{C} = \{(I_1^S, I_1^T), (I_2^S, I_2^T), \ldots, (I_\Omega^S, I_\Omega^T)\}$, where, depending on the corpus type, item pairs may be sentences, paragraphs, chunks, documents, etc. For parallel corpora, the item pairs are pairs of sentences. The goal is to extract translation pairs from the item-aligned set using only *internal distributional evidences*. Internal evidences, according to Kay and Röscheisen [143], represent information derived only from the given corpora themselves. Our criteria for establishing translational equivalence between two words are derived from this trivial case:

Imagine a scenario where a source word $w_1^S$ occurs only once on the source side of the corpus $\mathcal{C}$, in a source item $I_j^S$. There is a target word $w_2^T$ occurring in a target item $I_j^T$ (which is aligned to $I_j^S$) and the word $w_2^T$ also occurs only once on the target side of the corpus $\mathcal{C}$. Additionally, there does not exist another source word $w_a^S$ such that it occurs only once on the source side of the corpus and, at the same time, exactly in the item $I_j^S$, and there does not exist another target word $w_b^T$ that occurs only once on the target side of the corpus and exactly in the item $I_j^T$. Our key assumption is that the words $w_1^S$ and $w_2^T$ should then be listed as a translation pair. We can further generalize the intuition, that is, two words are extracted as a potential translation pair if they both satisfy the entire set of features $\mathcal{F}$, and there are no other words that

satisfy this set of features.[1]

The set $\mathcal{F}$ may include various clues as features, but in our work we opt only for the internal, language pair independent features that are related to the distributions of words over corpora. A source word $w_1^S$ and a target word $w_2^T$ are listed as a potential translation pair if they fulfil the following criteria:

1. The overall frequency of $w_1^S$ on the source side of the corpus is equal to the overall frequency of $w_2^T$ on the target side of the corpus.
2. The overall frequency of both words is above some minimum frequency threshold $F_f$.
3. $w_1^S$ and $w_2^T$ occur only in aligned item pairs, and with exactly the same frequency.
4. The number of aligned item pairs in which the words occur is above some minimum $F_i$.
5. There is no source language word $w_a^S \neq w_1^S$ such that the pair $(w_a^S, w_2^T)$ satisfies all the previous conditions, and there is no target language word $w_b^T \neq w_2^T$ such that the pair $(w_1^S, w_b^T)$ satisfies all the previous conditions.[1]

For instance, if the French word *pluie* occurs four times in the whole corpus, twice in item $I_j^S$, once in item $I_k^S$, and once in item $I_l^S$, and there is the English word *rain* that also occurs four times in total, twice in item $I_j^T$, once in item $I_k^T$, and once in item $I_l^T$, and there are no other words with the same frequency distribution in the corpus, we claim *(pluie, rain)* to be a potential translation pair.

In this chapter, we have opted for the listed criteria/constraints, but one is free to adjust or add more criteria if one desires to boost a certain behavior of the model, that is, if the focus should be more on accuracy or on coverage of the lexicon. By imposing, for instance, stricter thresholds for $F_f$ or $F_i$ (e.g., accepting only candidates that occur in at least two items), we can direct the algorithm for lexicon extraction towards higher accuracy, and, vice versa, by relaxing the thresholds, we boost the coverage of the lexicon.

[1] This specifies one-to-one alignment constraint, but more relaxed criteria are also possible. For instance, we could allow two or more target words to have the same features as a source word and then distribute partial link counts over all target candidates.

Finally, the proposed criteria for extraction of translation pairs are not biased towards high-frequency or low-frequency words, as they treat all words the same, trying to find potential translation pairs according to the defined set of features. However, in practice, due to the Zipf's law, the majority of the matched words in these pairs will be low-frequency words.

### 5.2.3 SampLEX: The Algorithm for Lexicon Extraction

By employing the aforementioned criteria for extraction of translation pairs on the initial corpus $\mathcal{C}$, we are able to extract only a limited number of translation pairs, since distributional evidences for the large corpus $\mathcal{C}$ are fixed and unchangeable. But by sampling data from $\mathcal{C}$, we actually build a new corpus, a sub-corpus $\mathcal{SC} \subset \mathcal{C}$ of size $\omega < \Omega$, which now has a changed set of distributional evidences, which may lead to extracting additional translation pairs. The process of data reduction may be observed as a process of *sampling*, that is, we randomly draw a subset of item pairs from $\mathcal{C}$, and build a new sub-corpus $\mathcal{SC}$. We can then repeat the process, sample another sub-corpus and try to detect more potential translation pairs.

Having the large corpus $\mathcal{C}$ of a finite size $\Omega$, the number of different sub-corpora is huge, but finite. The exact number of different sub-corpora that can be sampled from $\mathcal{C}$ is $\sum_{\omega=1}^{\Omega} \binom{\Omega}{\omega}$. Since we are clearly unable to process all the possible sub-corpora, we need to design a smart strategy to: (1) cover the entire initial corpus and (2) detect translation pairs for both high-frequency and low-frequency words.

**One Sampling Round with Fixed Sub-Corpora Size.** Let us fix the size of sub-corpora to some value $\omega$. We want to assure that every item pair from $\mathcal{C}$ is taken into account in at least one sub-corpus of size $\omega$. Additionally, we want to be able to repeat the procedure and obtain more different sub-corpora of the same size. The procedure of splitting the corpus $\mathcal{C}$ into sub-corpora of size $\omega$ which satisfies these constraints is summarized in alg. 5.1.

We build a set of $\lfloor \frac{\Omega}{\omega} \rfloor - 1$ sub-corpora of size $\omega$ and one sub-corpus of size $\omega + \Omega$ mod $\omega$, while, at the same time, we ensure that the complete original corpus $\mathcal{C}$ is covered.[2] We will call the described procedure the *sampling round*. If we want to repeat the procedure and acquire another set of sub-corpora of the same size, we simply go back to step 2 of the procedure in alg. 5.1 and perform another sampling round.

[2]Other options when dealing with the remainder of $\Omega$ mod $\omega$ sentences are: (1) simply leave the remainder out of any calculations, or (2) make another, $(\lfloor \frac{\Omega}{\omega} \rfloor + 1)$-th sub-corpus which encompasses these $\Omega$ mod $\omega$ sentences. However, the choice of the heuristic does not have any influence on the overall results in the task of bilingual lexicon extraction.

**Algorithm 5.1**: ONE SAMPLING ROUND WITH FIXED $\omega$

**Input** : Corpus $\mathcal{C}$ of size $\Omega$; Fixed sub-corpus size $\omega$ ;
1: **detect** the number of sub-corpora for this sampling round: $\lfloor \frac{\Omega}{\omega} \rfloor$ ;
2: **shuffle** randomly the item pairs in $\mathcal{C}$ to obtain a permutation of the item pairs in $\mathcal{C}$ ;
3: **split** $\mathcal{C}$ into sub-corpora of equal size $\omega$ as follows:
**for** $i \leftarrow 1$ **to** $\lfloor \frac{\Omega}{\omega} \rfloor - 1$ **do**
  (a): **assign** the item pairs from position $(i-1) \cdot \omega + 1$ until position $i \cdot \omega$ to the sub-corpus $\mathcal{SC}_i$ ;
(b): **assign** the remaining item pairs from position $(\lfloor \frac{\Omega}{\omega} \rfloor - 1) \cdot \omega + 1$ until the end (position $\Omega$) to the sub-corpus $\mathcal{SC}_{\lfloor \frac{\Omega}{\omega} \rfloor}$ ;
**Output**: A division of $\mathcal{C}$ into $\lfloor \frac{\Omega}{\omega} \rfloor$ sub-corpora ;

**The Final Algorithm: SampLEX**. Now, we have everything set for the construction of the algorithm. In order to capture words with different frequencies, we need to vary the sub-corpora size $\omega$. With respect to the Zipf's law [243], we have decided to vary the values of $\omega$ from $\Omega$ down to 1, where $\omega$ is divided by 2 in each step of the loop. In that way, we ensure that all the words occur as low-frequency words in at least some sub-corpora of various sizes. Again, if we want to reduce frequencies of high-frequency words, we need samples of smaller sizes, so such words will typically learn its candidate translations from sampled sub-corpora consisting of only a few sentences. One pass of the algorithm from the values $\Omega$ to 1 is called an *iteration*.

We can detect potential translation pairs in many different sub-corpora (of various sizes). Additionally, we should assign more weight to translation pairs that fulfil the strict criteria in sub-corpora of larger size $\omega$. For instance, if we detect that two words have identical frequency distributions and have fulfilled all the criteria from sect. 5.2.2 in a sub-corpus consisting of a few million items, that evidence should be more important than detecting that the two words could be extracted from a sub-corpus comprising only a few sentences. Thus, for each potential translation pair $t_{ij}$ we assign a corresponding overall score $os_{t_{ij}}$. If we detect that the two words that form the potential translation pair $t_{ij}$ could be extracted from a sub-corpus of size $\omega$, we update the score $os_{t_{ij}} := os_{t_{ij}} + 1 \cdot weight_\omega$, where $weight_\omega = \lfloor \frac{\Omega}{\omega} \rfloor$. This way we assign more importance when the pairs are extracted from larger sub-corpora. For instance, if we detect that two words from the potential translation pair $t_{ij}$ are extracted from the original corpus $\mathcal{C}$, then $\omega = \Omega$ and $os_{t_{ij}} := os_{t_{ij}} + 1$.

The final algorithm is presented in alg. 5.2. We will call this procedure the *SampLEX* (SAMPling + LEXicon) algorithm. The proposed algorithm exhibits

**Algorithm 5.2**: SAMPLEX ALGORITHM

**Input** : Initial large corpus $\mathcal{C}$ of size $\Omega$ ;

1: **initialize**:

(a): **define** the criteria for extraction of potential translation pairs from sub-corpora (see sect. 5.2.2) ;

(b): **initialize** an empty lexicon $L_f$ (each entry in the lexicon $L_f$ will have the following form: $(t_{ij}, os_{t_{ij}})$, where $t_{ij}$ denotes the extracted potential translation pair consisting of a source language word $w_i^S$ and a target language word $w_j^T$, while $os_{t_{ij}}$ is a variable that denotes the overall score for the potential translation pair $t_{ij}$) ;

2: **set** initial sub-corpora size: $\omega := \Omega$ ;

3: **perform** one *sampling round* with the current sub-corpora size set to $\omega$ (see sect. 5.1) $\rightarrow$ we obtain $\lfloor \frac{\Omega}{\omega} \rfloor$ different sub-corpora: $\mathcal{SC}_1, \ldots, \mathcal{SC}_{\lfloor \frac{\Omega}{\omega} \rfloor}$, all of size $\omega$ except the last one (its size is always $\omega + \Omega \bmod \omega$) ;

4: **extract** potential translation pairs from all sub-corpora obtained in step 3.

5: **if** a potential translation pair $t_{ij}$ is already present in the lexicon $L_f$ **then**
  **update** the score $os_{t_{ij}} := os_{t_{ij}} + 1 \cdot weight_\omega$ ;
**else**
  **add** $t_{ij}$ to $L_f$ ;
  **set** $os_{t_{ij}} := 1 \cdot weight_\omega$ ;

6: **set** new sub-corpora size: $\omega := \lfloor \omega/2 \rfloor$ ;

7: **if** $\omega > 0$ **then**
  **go** to step 3 ;
**else**
  (end of one *iteration*) ;
  **go** to step 8 ;

8: **check** the stopping criteria:

**if** no new translation pairs were extracted after the end of one whole iteration **or** we have reached the maximum or the predefined number of iterations **or** timeout **then**
  **end** $\rightarrow$ **output** the lexicon $L_f$ ;
**else**
  **go** to step 2 ;

**Output**: The final lexicon $L_f$

only one possible strategy for mining translation pairs from sub-corpora. For instance, we could opt for another strategy when deciding how to change the size of sub-corpora (step 6), skip already processed sub-corpora (step 4), remodel the criteria for extraction from sect. 5.2.2 (step 1(a)), change stopping criteria (step 8), or employ a procedure for the sub-corpora sampling different from the one presented in alg. 5.1 (step 3). However, our main goal is to propose a general framework for lexicon extraction when the data sampling approach is employed, where other researchers could design their own algorithms supported by the same idea.

### 5.2.4 Properties of the SampLEX Algorithm

Reducing corpora size provides several benefits. First, establishing associations between pairs of words is much easier when we deal with low-frequency words - we reduce our problem to a *binary decision problem*. According to the specified criteria for extraction, two words are simply considered to be a potential translation pair, or they are not. By employing the criteria that rely on raw frequency counts as distributional evidences, we remove the need of an association measure based on hypothesis testing such as the $G^2$ statistic [82, 4] or a similarity-based measure such as the Dice coefficient [74], which are often unreliable when dealing with low-frequency words [186].

The SampLEX algorithm is *symmetric* and *non-directional*. The final output of the algorithm provides translation pairs along with their corresponding scores obtained after training. We can easily transform these scores into word translation probabilities to build word translation tables similar to those of IBM Model 1. Since the algorithm is symmetric, we can obtain both source-to-target and target-to-source word translation probabilities after the algorithm run is completed:

$$P(w_2^T|w_1^S) = \frac{os_{t_{12}}}{\sum_j os_{t_{1j}}} \qquad P(w_1^S|w_2^T) = \frac{os_{t_{12}}}{\sum_j os_{t_{j2}}} \tag{5.1}$$

Surprisingly, another modeling advantage lies in *randomness* when selecting sub-corpora. Namely, if we detect that two words constantly co-occur in aligned items randomly sampled from the large corpus, regardless of the surrounding context, it actually strengthens the confidence that those two words really constitute a translation pair. During the sampling procedure, sentences are moved from their "natural" surrounding of other sentences (the context in this case) and the new sub-corpus is built by randomly taking sentences from the entire corpus.

## 5.3 State-of-the-Art Models for BLE

In order to evaluate the performance of our SampLEX algorithm for bilingual lexicon extraction, we compare it with other models that constitute state-of-the-art for BLE from parallel data, and are often used in real-life applications (see sect. 5.1). We briefly introduce three representative BLE models which we use in later evaluations and comparisons: (1) IBM Model 1 (sect. 5.3.1), (2) the model relying on the Dice coefficient association measure (sect. 5.3.2), and (3) the model relying on the log-likelihood ratio (sect. 5.3.3).

### 5.3.1 IBM Model 1

The first state-of-the-art BLE model is IBM Model 1 for word alignment [37, 229]. This word alignment model is a purely lexical model, that is, the only set of parameters employed by the model are word translation probabilities. Since word alignment models are not the focus of this work, we omit the exact generative story for IBM Model 1, but the curious reader may find all the details in [37] or [229]. Word translation probability $P(w_2^T|w_1^S)$ denotes a probability that a source language word $w_1^S$ generates a target language word $w_2^T$. These probabilities can then be used to decide upon translational equivalence between words and to build bilingual lexicons from parallel texts.[3] That makes it comparable to our SampLEX algorithm, which can also output word translation probabilities (see eq. (5.1)). IBM Model 1 is used in many systems as a primary tool for bilingual lexicon extraction from parallel data (e.g., [308, 214, 215, 171]).

### 5.3.2 DICE Model

Another BLE model is a similarity-based model relying on the Dice coefficient (DICE):

$$DICE(w_1^S, w_2^T) = \frac{2 \cdot C(w_1^S, w_2^T)}{C(w_1^S) + C(w_2^T)} \tag{5.2}$$

where $C(w_1^S, w_2^T)$ denotes the co-occurrence count of words $w_1^S$ and $w_2^T$ in the aligned items from the corpus. $C(w_1^S)$ and $C(w_2^T)$ denote the count of $w_1^S$ on the source side of the corpus, and the count of $w_2^T$ on the target side of the corpus, respectively. The Dice coefficient was used as an associative method for word alignment by Och and Ney [229]. Tiedemann [289] used it as one

[3] We have also tried to use word translation probabilities from the higher order IBM Models 2-6, but we have not detected any major difference in results of the BLE task.

| | $w_1^S$ | $\neg w_1^S$ |
|---|---|---|
| $w_2^T$ | $k$ | $l$ |
| $\neg w_2^T$ | $m$ | $n$ |

Table 5.1: The contingency table for a pair of words $(w_1^S, w_2^T)$.

associative clue for his clue-based word alignment, and Melamed [197] used it to measure the strength of translational equivalence.

### 5.3.3 LLR Model

Another associative model that we compare against is based on the log-likelihood-ratio (LLR), which is derived from the $G^2$ statistic [82]. LLR is a more appropriate hypothesis testing method for detecting word associations from limited data than the $\chi^2$ test [186] and was previously used as an effective tool for automatically constructing bilingual lexicons [197, 209, 215]. Its definition is easily explained on the basis of a contingency table [147, 231], which is a four-cell matrix for each pair of words $(w_1^S, w_2^T)$ (see tab. 5.1).

The contingency table records that source word $w_1^S$ and target word $w_2^T$ co-occur in $k$ aligned item/sentences pairs, and $w_1^S$ occurs in $m$ aligned pairs in which $w_2^T$ is not present. Similarly, $w_2^T$ occurs in $l$ aligned pairs in which $w_1^S$ is not present, and $n$ is the number of aligned pairs that involve neither $w_1^S$ nor $w_2^T$. The final formula for the log-likelihood ratio is then defined as:

$$\begin{aligned} LLR(w_1^S, w_2^T) = G^2(k,l,m,n) = 2(&k\log k + l\log l + m\log m + n\log n \\ &- (k+l)\log(k+l) - (k+m)\log(k+m) \\ &- (l+n)\log(l+n) - (m+n)\log(m+n) \\ &+ (k+l+m+n)\log(k+l+m+n)) \end{aligned} \quad (5.3)$$

High LLR scores can indicate either a positive association or a negative one [212]. Since we expect translation pairs to be positively associated, we impose an additional constraint: $P(w_1^S, w_2^T) > P(w_1^S) \cdot P(w_2^T)$, where $P(w_1^S, w_2^T) = \frac{k}{k+l+m+n}$, $P(w_1^S) = \frac{k+m}{k+l+m+n}$ and $P(w_2^T) = \frac{k+l}{k+l+m+n}$. This constraint retains only positively associated words as potential translation pairs.

## 5.4 Experimental Setup

**Training Collections.** We work with parallel Europarl data [153] (see sect. 3.2.1 in chapter 3) for Dutch-English and Italian-English language pairs,

retrieved from the website[4] of the OPUS project [290]. We use subsets of the corpora, comprising the first 300,000 sentence pairs. For Dutch-English, there are 76,762 unique Dutch words, and 37,138 unique English words. For Italian-English, there are 68,710 unique Italian words and 37,391 unique English words. The unbalance between the number of unique vocabulary words is mostly due to a richer morphological system in Italian and the noun compounding phenomenon in Dutch.

Since we also want to test and evaluate the behavior of our system in a setting where only limited parallel data are present, we construct additional subsets of the Europarl data comprising only the first 2,000, 10,000 and 50,000 sentence pairs from the corpora.

**Training Setup.** Parameter values are set to the same values for all training datasets. We set $F_f = F_i = 0$, which means that all words that occur in a sub-corpus at least once may be extracted. By setting some higher thresholds $F_f$ and $F_i$, we could move the algorithm towards extracting lexicons of higher accuracy, but lower coverage. Unless noted otherwise, we stop our training procedure for SampLEX after 1,000 iterations for all corpora. The SampLEX algorithm converges quickly - many translations are found in the first few iterations. However, having more iterations implies obtaining more different evidences from different sub-corpora and assigning more significance for the extracted candidates (see sect. 5.2.4). Therefore, we have decided to use 1,000 iterations for safety. The analysis of the convergence of our SampLEX algorithm will be provided in more detail later (see sect. 5.5.4). Other stopping criteria for the SampLEX algorithm are also possible (see step 8 in alg. 5.2).

For IBM Model 1, we use standard GIZA++ settings and train IBM Model 1 with 5 iterations (IBM1-i5) and 20 iterations (IBM1-i20) of the expectation-maximization (EM) algorithm [70, 37], as often found in the literature [229, 211].

**Ground Truth Translation Pairs.** In order to evaluate the BLE models, we have designed a set of ground truth translation pairs - we have randomly sampled a set of Dutch words that occur in the full corpus comprising 300, 000 sentences. Following that, we have used the *Google Translate* tool plus an additional annotator to translate those words to English. The annotator has manually revised the lists and retained only words that have their corresponding translation in the English vocabulary. In order to build a one-to-one ground truth dataset of translation pairs, only one possible translation has been annotated as correct. In case when more than one translation is possible, the annotator has marked as correct the translation that occurs more frequently in the English Europarl data. Finally, we have obtained a set of 1,001 ground truth one-to-one

[4] `http://opus.lingfil.uu.se/Europarl3.php`

translation pairs. We have followed the same procedure for Italian-English and have also constructed a set of 1,001 ground truth translation pairs.

**Evaluation Metrics.** Let us retain only the best scoring candidate translation for each word from the obtained lists of potential translation pairs, and build a non-probabilistic lexicon of one-to-one word translations: $L_e$. Assuming that we now have a set $G$ of ground truth one-to-one word translation pairs, we can evaluate the quality of our lexicon with respect to the ground truth set $G$. We use standard precision, recall and equally balanced F-measure ($\eta$=1, also called *F-1 score*) [306] as our evaluation metrics:

$$Prec_{L_e,G} = \frac{|L_e \cap G|}{|L_e|} \qquad Rec_{L_e,G} = \frac{|L_e \cap G|}{|G|} \tag{5.4}$$

$$F_{L_e,G} = (1+\eta^2)\frac{Prec_{L_e,G} \cdot Rec_{L_e,G}}{\eta^2 \cdot Prec_{L_e,G} + Rec_{L_e,G}} \tag{5.5}$$

Since sometimes a word has more than one correct translation (e.g., Dutch word *verklaring* can be translated as *statement*, *declaration* or *explanation*), and the current evaluation setting cannot capture that phenomenon, we also evaluate the quality of the lexicon in a more lenient setting, where, instead of performing the hard cut-off, that is, instead of retaining only the top candidate translation for a word, we retain the list of all candidate translations with their corresponding scores and calculate the *mean reciprocal rank* (MRR) [309]. For a source language word $w_1^S$, $rank(GTC(w_1^S))$ denotes the rank of its correct translation (as provided by the set of ground truth translation pairs) within the retrieved list of candidate translations. MRR of the lexicon is then defined by the following formula:

$$MRR_{L_e,G} = \frac{1}{|L_e|} \sum_{w_1^S \in L_e} \frac{1}{rank(GTC(w_1^S))} \tag{5.6}$$

All scores reported in all experiments in the following section are obtained as an average over five independent runs of the SampLEX algorithm.

## 5.5 Experiments, Results and Discussion

We conduct several experiments to measure the properties of the SampLEX algorithm and the quality of the lexicon constructed using the algorithm: (1) We evaluate the lexicon obtained by SampLEX using the full corpus of 300,000 sentences, and compare its accuracy with the accuracy of state-of-the-art systems

| | Dutch-English | | | | |
|---|---|---|---|---|---|
| | **IBM1-i5** | **IBM1-i20** | **DICE** | **LLR** | **SampLEX** |
| Prec(300k) | 0.711 | 0.702 | 0.696 | 0.766 | **0.822** |
| MRR(300k) | 0.820 | 0.805 | 0.777 | 0.854 | **0.907** |

Table 5.2: Precision and MRR scores for all models trained on the first 300,000 sentences of Dutch-English Europarl data, and evaluated on the sets of 1,001 ground truth translation pairs for Dutch-English.

| | Italian-English | | | | |
|---|---|---|---|---|---|
| | **IBM1-i5** | **IBM1-i20** | **DICE** | **LLR** | **SampLEX** |
| Prec(300k) | 0.791 | 0.775 | 0.793 | 0.836 | **0.877** |
| MRR(300k) | 0.878 | 0.859 | 0.849 | 0.895 | **0.925** |

Table 5.3: Precision and MRR scores for all models trained on the first 300,000 sentences of Italian-English Europarl data, and evaluated on the sets of 1,001 ground truth translation pairs for Italian-English. All models (including SampLEX) provide translations for all 1,001 from the ground truth test set.

from sect. 5.3 trained on the same corpus; (2) After performing the error analysis, we carry out another set of experiments that prove that SampLEX, due to its modeling properties, alleviates the problem of indirect associations; (3) We test our lexicon in a setting where only limited parallel data are available and show that the SampLEX-based lexicon outperforms other bilingual word lexicons in that setting in terms of quality provided by the F-measure and precision scores; and finally (4) We investigate the convergence properties of SampLEX.

### 5.5.1 Experiment I: Testing the Quality of the SampLEX Lexicon in Terms of Precision

Unlike our baseline state-of-the-art systems for BLE, the SampLEX algorithm does not assure the full coverage of the source vocabulary, as it does not necessarily build ranked lists of candidate translations for all the words observed during training. However, our claim is that translation pairs obtained by SampLEX are of higher quality than those obtained by the baseline systems. Therefore, in line with research question RQ3, with this experiment we want to answer the following question: "Are translation pairs obtained by the SampLEX algorithm really more reliable than translation pairs obtained by other methods?". In order to answer that question, we calculate precision and MRR scores on our ground truth datasets for Italian-English and Dutch-English, where all the BLE

models have been trained on the full 300,000 sentences datasets. The obtained scores are presented in tab. 5.2 and tab. 5.3.

As previously shown by Moore [211], LLR serves as a better associative method than the Dice coefficient for the word alignment task. We obtain the same finding for bilingual lexicon extraction. Additionally, the model based on LLR is also better than IBM Model 1 when applied for BLE. Munteanu and Marcu [215] drew the same conclusion, and they used the LLR-based lexicon in their system when a higher precision of the lexicon was paramount. However, the results reveal that the quality of the lexicon obtained by the SampLEX algorithm is superior to the LLR-lexicon *in terms of precision* and, consequently, to all other evaluated lexicons. Furthermore, since SampLEX finds translations for all 1,001 from the ground truth test set, we may safely claim that the SampLEX algorithm does not suffer from low recall, that is, it does not trade high precision scores for low recall scores.

### 5.5.2 Experiment II: Investigating Indirect Associations

When examining the results, we have detected that one advantage of our SampLEX algorithm is due to its mitigation of the phenomenon of the so-called indirect associations. Indirect associations, as defined by [197], are associations between words that have a tendency to co-occur much more often than expected by chance, but are not mutual translations. Lexicon extraction models unaware of the indirect associations tend to give translational preference to higher-frequency words. Considering the fact that one key assumption of our model is sub-corpora sampling that causes decreasing frequencies of words in the obtained sub-corpora from which translation pairs are learned, our model should successfully mitigate the problem of indirect associations.

Indeed, during the error analysis, we have detected that both IBM Model 1 and LLR provide an incorrect translation of the Dutch word *beschouwen (consider)*, since both models retrieve the English word *as* as the first candidate translation (due to a very high frequency of the collocation *consider as*). Other examples of the same type include the Dutch word *integreren (integrate)* which is translated as *into*, *betwijfelen (doubt)* which is translated as *whether*, or an Italian example of the verb *entrare (enter)* which is translated as *into*. Our BLE model, on the other hand, provides correct translations for all these examples.

Additionally, in [57] it was already noted that collocates often tend to cause confusion among algorithms for bilingual lexicon extraction. More examples include the Dutch word *opinie (opinion)*, translated as *public* by IBM Model 1 and LLR (due to a high frequency of the collocation *public opinion*), the Dutch word *cirkels (circles)*, translated as *concentric*, or the Italian word

*pensionabile (pensionable)*, translated as *age*. All these examples are again correctly translated by our model for lexicon extraction.

In order to test the hypothesis that our lexicon extraction model does not suffer from the problem of learning indirect associations, we have conducted a small experiment. For the purpose of the evaluation, we have constructed a small dataset of 219 Italian verbs in first person plural of the present tense. We have also constructed the set of ground truth translations in the same way as in sect. 5.4. These verbs are easy to extract because they all have the same regular suffix *-iamo* (e.g., the verb *respiriamo*, meaning *(we) breathe*). If the problem of indirect associations for a lexicon extraction method is prominent, the English word *we* will appear as the first translation for many of these verbs, instead of the word that really bears the content of the verb (e.g., *breathe*). Tab. 5.4 displays precision and MRR scores for the lexicon extraction models evaluated on this toy dataset.

In addition, we believe that by introducing additional knowledge (e.g., POS tags) and restricting translations to, e.g., retain the same POS tag as the original source word (e.g., an English noun has to be translated as a noun in Italian or Dutch), we may further improve the quality of the lexicons induced by any of these models (see also later chapter 6).

| | **IBM1-i5** | **DICE** | **LLR** | **SampLEX** |
|---|---|---|---|---|
| Prec(300k) | 0.448 | 0.420 | 0.612 | **0.858** |
| MRR(300k) | 0.558 | 0.511 | 0.730 | **0.914** |

Table 5.4: Precision and MRR scores on our evaluation set consisting of Italian *-iamo* verbs (present tense, first person plural).

As expected, due to its modeling property related to the reduction of word frequencies, our model of BLE from parallel data does not suffer from the problem of indirect associations like other models. That property eventually has a positive impact on precision and MRR scores and the overall lexicon quality.

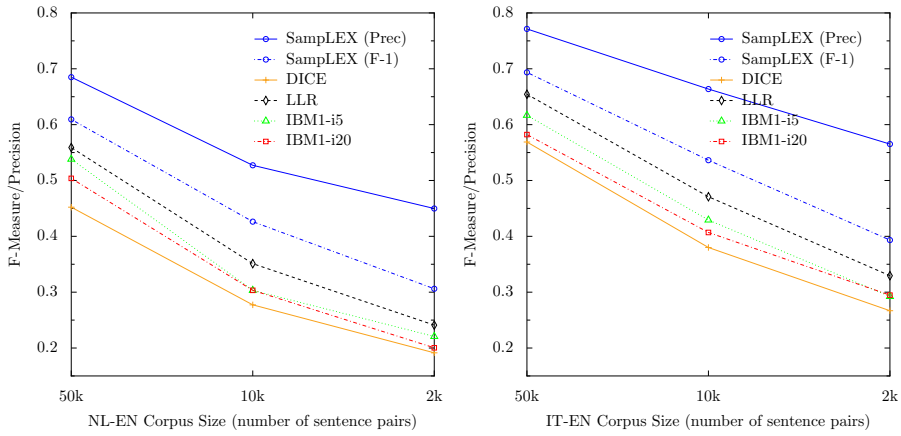
### 5.5.3 Experiment III: Experiments with a Limited Amount of Parallel Data

In a real-life situation, one often possesses only limited parallel data (e.g., terminology texts from specific, very narrow domains and sub-domains). With this set of experiments we test the performance of all our BLE models in comparison in such a setting with limited parallel data. To simulate the shortage of data, we have extracted three additional corpora of smaller sizes by selecting the first $2,000$, $10,000$ and $50,000$ sentence pairs from our Dutch-English and Italian-English Europarl data. From our initial ground truth set (see sect. 5.4),

we have only retained words that occur at least once in the respective corpora as the ground truth for evaluations (i.e., there are 444 words in the ground truth dataset for the corpus consisting of the first Dutch-English 2,000 sentence pairs, 768 for the corpus comprising the first 10,000 sentence pairs, and 931 words for the corpus consisting of the first 50,000 Dutch-English sentence pairs). Our research question is now: "Are lexicons extracted by SampLEX really of better quality than lexicons obtained by other methods when dealing with parallel corpora of limited size?"

As mentioned before, the SampLEX algorithm does not have a property to provide candidate translations for the entire source language vocabulary, but we claim that SampLEX is directed towards extracting only highly reliable and precise potential translation pairs which, consequently, leads to higher-quality bilingual lexicons. That claim is again supported by the findings presented in fig. 5.1a for Dutch-English, and in fig. 5.1b for Italian-English. Since SampLEX does not necessarily obtain the lists of translations for all words in a vocabulary, its precision scores are different than its F-measure scores. For all other models within this evaluation setting, it is valid: *Precision=Recall=F-1 score*.

We have also performed an additional experiment to test whether the candidate translations for Dutch and Italian words that happen to be retrieved by the SampLEX algorithm still display better overall precision and MRR scores than the candidate translations for the same Dutch and Italian words obtained by the other methods. If that is not true, we could use SampLEX only to extract source words for which a translation might be found, but the particular translation for




(a) Dutch to English (b) Italian to English

Figure 5.1: Precision and F-1 scores over (a) Dutch-English, and (b) Italian-English parallel corpora of different size (2k, 10k, 50k sentence pairs).

| | Dutch-English | | | | |
|---|---|---|---|---|---|
| | **IBM1-i5** | **IBM1-i20** | **DICE** | **LLR** | **SampLEX** |
| Prec(2k) | 0.367 | 0.362 | 0.332 | 0.432 | **0.450** |
| MRR(2k) | 0.421 | 0.420 | 0.397 | 0.450 | **0.484** |
| Prec(10k) | 0.427 | 0.431 | 0.368 | 0.489 | **0.527** |
| MRR(10k) | 0.507 | 0.504 | 0.451 | 0.559 | **0.585** |
| Prec(50k) | 0.618 | 0.595 | 0.530 | 0.662 | **0.685** |
| MRR(50k) | 0.707 | 0.690 | 0.618 | 0.718 | **0.743** |

Table 5.5: Precision and MRR scores on Dutch-English for all models trained on the subsets of different sizes (2k, 10k, 50k sentences).

| | Italian-English | | | | |
|---|---|---|---|---|---|
| | **IBM1-i5** | **IBM1-i20** | **DICE** | **LLR** | **SampLEX** |
| Prec(2k) | 0.509 | 0.517 | 0.435 | 0.552 | **0.565** |
| MRR(2k) | 0.580 | 0.578 | 0.490 | **0.611** | 0.608 |
| Prec(10k) | 0.598 | 0.585 | 0.501 | 0.646 | **0.664** |
| MRR(10k) | 0.656 | 0.649 | 0.571 | 0.691 | **0.701** |
| Prec(50k) | 0.713 | 0.697 | 0.638 | 0.758 | **0.771** |
| MRR(50k) | 0.793 | 0.785 | 0.719 | 0.806 | **0.828** |

Table 5.6: Precision and MRR scores on Italian-English for all models trained on the subsets of different sizes (2k, 10k, 50k sentences).

each extracted word could then be obtained by some other method. However, it is not the case, as the results in tab. 5.5 and tab. 5.6 reveal. As noted in the literature [186], we observe that, out of all baseline models for BLE, LLR suffers the least from data sparsity, but still performs worse than our method.

Since SampLEX is supported by the data sampling paradigm, the criteria for extracting candidate translations and the whole training process inherently remain the same when working with parallel corpora of limited size. However, it is natural that the results decrease when the size of the large corpus $\mathcal{C}$ decreases. The more data we possess, the more sub-corpora we can sample, which finally provides better chances to extract correct translation pairs. We could say that SampLEX takes the best of both worlds - it benefits from the idea of data reduction, yet it provides better scores when more input data are available.

Our results reveal that there is still room for improvement. For instance, using the criteria from sect. 5.2.2, we have detected that our algorithm fails on fixed collocations and phrases if they are not processed as single expressions

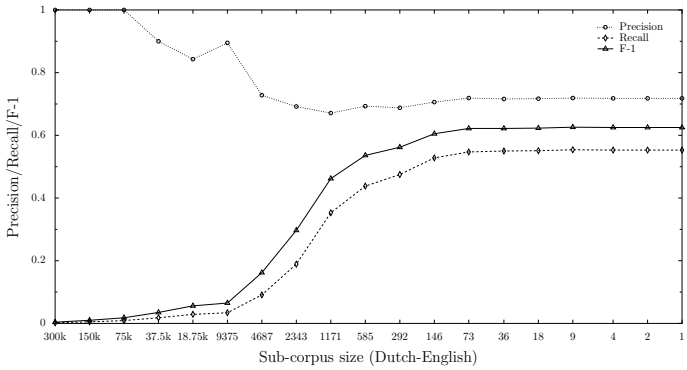



Figure 5.2: Precision, recall and F-1 scores for Dutch-English over the sequence of sampling rounds in the first iteration of the SampLEX algorithm.

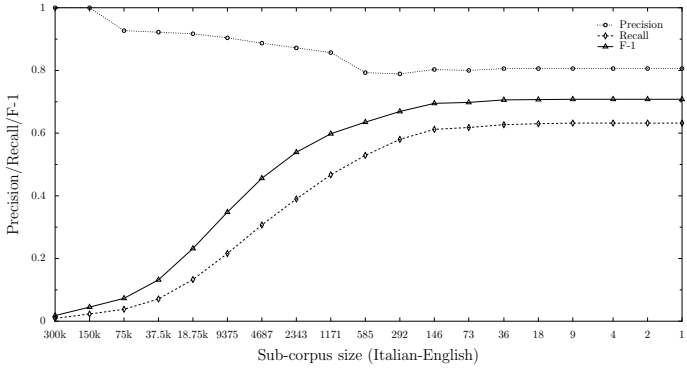
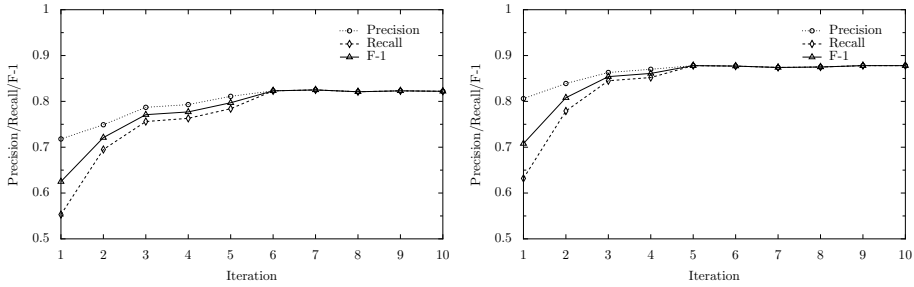



Figure 5.3: Precision, recall and F-1 scores for Italian-English over the sequence of sampling rounds in the first iteration of the SampLEX algorithm.

beforehand (see also appendix A). Similar to alignment models [37, 229], high-frequency words (such as stop words) are another issue, due to their high ambiguity and their tendency to generate links to unrelated terms simply on the basis of their high frequency in the corpus.

### 5.5.4 Experiment IV: Investigating Convergence

With the final set of experiments, we aim to test the rate of convergence and convergence properties of the SampLEX algorithm. We again operate with the full collection comprising 300,000 sentence pairs and the same test collections of 1,001 Italian and Dutch words. First, we track recall and precision scores *during sampling rounds in the first iteration*. Fig. 5.2 displays the results of the test for Dutch-English, while fig. 5.3 displays the results for Italian-English. We observe the following:




(a) Dutch to English (b) Italian to English

Figure 5.4: Precision, recall and F-1 scores over the first 10 iterations of SampLEX for (a) Dutch-English, and (b) Italian-English.

(i) Potential translation pairs that are extracted from larger sub-corpora are highly reliable, but rare. This phenomenon is reflected in very high precision, but very low recall scores, and it is also in line with the intuition which states that it is more difficult to fulfil the criteria for extraction of translation pairs (see 5.2.2) in a large sub-corpus, and therefore the extracted pairs should be rare and highly reliable. Smaller sub-corpora size leads to more relaxed criteria for extraction which consequently leads to extracting more translation pairs that become less and less reliable as the sub-corpus size decreases. This phenomenon is reflected in recall scores that continuously increase over the sequence of rounds in an iteration, while precision scores decrease. However, the algorithm exhibits a desirable property that the decrease of precision occurs at the slower rate than the increase of recall.

(ii) Potential translation pairs that are extracted from very small sub-corpora (50 sentences or less) do not have a high impact on overall recall and precision scores. This phenomenon is again in concordance with the Zipf's law. Namely, pairs extracted from such small sub-corpora are typically between high-frequency words (see sect. 5.2.3). According to the Zipf's law, there is a limited small set of high-frequency words in the entire corpus (here, we are talking about the original large corpus), hence the extracted translation pairs also cover a limited small set of potential pairs, and the algorithm does not introduce plenty of new knowledge to the lexicon. Consequently, it leads to only minor changes in overall recall and precision scores, as visible in fig. 5.2 and fig. 5.3.

In the next experiment, we record recall and precision scores *in the subsequent iterations of the algorithm*. The results are provided in fig. 5.4a (Dutch-English) and fig. 5.4b (Italian-English). The obtained results again reveal several interesting phenomena:

(iii) SampLEX converges quickly, and stable recall and precision levels are

obtained after only a few iterations for both language pairs. We can also observe a continuous gradual increase in recall and precision scores (and, consequently, in F-1 scores) in each iteration before the convergence is reached.

(iv) By observing more and more different sub-corpora, the algorithm smooths out the effect of rare phenomena that might by chance occur in the first stages of the algorithm and negatively affect precision. When a translational equivalence between two words is encountered in more different sub-corpora, it raises the significance of the event (see sect. 5.2.4). In short, by seeing more sub-corpora, the algorithm retains statistically significant equivalences and removes plenty of noise coming from the accidental rare equivalences in the first stages. Recall scores increase over iterations due to the fact that the algorithm, again by observing more different sub-corpora, learns more new translational equivalences that were previously not present in the lexicon.

## 5.6 Conclusions and Future Work

In part II we have tackled the problem of finding term translations in parallel corpora. We have tried to find an answer to research question RQ2, that is, how to automatically induce bilingual lexicons from multilingual parallel data without any other source of knowledge besides the data itself. These bilingual lexicons should include only reliable translation pairs if possible (research question RQ3).

As a major contribution, we have proposed a new statistical framework for the automatic construction of bilingual word lexicons from parallel corpora built upon the idea of sampling many smaller sub-corpora from an initial larger item-aligned corpus. The new SampLEX algorithm for bilingual lexicon extraction presented in this chapter is directed towards extraction of highly reliable word translation pairs. After comparisons with other models for BLE from parallel data, we have proven that SampLEX builds lexicons of higher quality as revealed by the F-measure and precision scores, which is especially important in a setting where only a limited amount of parallel data is available. The proposed framework allows for many further experiments (see sect. 5.2.3) and possible applications. We present one application and the utility of the proposed algorithm in the CAT tool (see appendix A) for legal translations. Another interesting application would be testing the automatically acquired bilingual lexicons as features in larger systems such as end-to-end statistical machine translation systems or cross-lingual search engines, or use the highly-reliable translation pairs provided by SampLEX as an additional source of knowledge in alignment algorithms or as seeds for some other BLE model.

In a broader perspective, the utility of the data reduction paradigm has proven its value in other tasks such as text classification or clustering (e.g., [226, 55]). We believe that the same paradigm may find its application in other NLP tasks beyond bilingual lexicon extraction from parallel data. Following that line of thinking, we have also tried to apply a similar sub-corpora sampling principle when tackling the task of bilingual lexicon extraction from comparable data. However, the application of this principle has not yielded any improvements in the setting where only a comparable corpus is available. Part III therefore introduces and describes a completely different approach to inducing bilingual lexicons from comparable data which relies on the knowledge of latent cross-lingual topics induced from such multilingual non-parallel data.

## 5.7 Related Publications

[1] **I. Vulić** and M.-F. Moens. "Sub-corpora sampling with an application to bilingual lexicon extraction," in *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India, 8-15 December 2012, pp. 2721-2738, ACL, 2012.

# Part III

# Modeling Cross-Lingual Semantic Similarity

# Outline of Part III

In part III, we make a transition from high-quality sentence-aligned parallel data which we utilized throughout part II to comparable corpora. Comparable corpora possess a much higher degree of noise, uncertainty and often exhibit a lack of structure. Altogether it means that approaches described in part II which were tailored for parallel corpora are not functional within this more difficult noisier setting. Therefore, one needs to find another way to tackle the problem of modeling similarity between words and phrases across languages and, consequently, inducing translation pairs in this more difficult and noisier setting (see sect. 3.2.2 in chapter 3). In other words, as already hinted in RQ5, we need to design a completely different algorithmic strategy in order to deal with comparable data.

Part III constitutes the heart of this thesis and makes a series of contributions to the field of *distributional semantics*. In this part, in order to provide answers to research questions RQ2, RQ3, and RQ5, we construct and describe a complete statistical framework for modeling *semantic similarity across languages*. The entire framework and the whole work reported in this part leans on the knowledge of latent cross-lingual topics/concepts which can be effectively induced from non-parallel data (see sect. 4.2.1 and sect. 4.3). The usage of these latent cross-lingual topics in distributional semantics has been pioneered in this thesis. The proposed framework is fully corpus-based and data-driven, as it does not rely on any other resource except for the given corpora. Additionally, the framework is purely statistical and it does not make any language pair specific assumptions. It should make the framework practically applicable and portable to plenty of language pairs (with more or less success).

In this part, we report a large body of research conducted within the scope of this thesis. We start from the fundamental new models of cross-lingual semantic similarity which rely on the knowledge of latent cross-lingual topics and gradually extend the framework with more complex models and other insights. Part III is logically divided into five standalone chapters, where each chapter covers one research theme within this framework, and has been published as a standalone research article. Consequently, each chapter provides one important major contribution to the whole framework as follows:

I. We introduce the framework for modeling semantic similarity across languages and automatically extracting bilingual lexicons based on latent cross-lingual topics/concepts, and present a set of new models which operate directly in the latent cross-lingual semantic space spanned by these latent topics (chapter 6).

II. We present a new approach to selecting only highly confident translation pairs from the lists of semantically similar words (research question RQ3). The

precision-oriented selection algorithm is extremely important in the noisy setting with non-parallel data (chapter 7).

III. We introduce a new cross-lingual semantic space spanned by all vocabulary words in both languages. The models in that semantic space rely on the concept of semantic responding or free word association, and are proven to be more robust and more effective than the models which operate directly with latent cross-lingual topics (chapter 8).

IV. We propose, describe and analyze a new framework for bootstrapping cross-lingual vector spaces from non-parallel data. The models supported by this bootstrapping framework advance current state-of-the-art in the fully data-driven models of cross-lingual semantic similarity (chapter 9).

V. As opposed to context-insensitive models of similarity, we demonstrate how to build context-sensitive models of similarity within the same (probabilistic) framework (chapter 10).

In addition to these listed major contributions covering major research themes, each chapter also provides a series of minor contributions related to each research theme, which we list in their corresponding chapters. Moreover, in the first chapter of this part (chapter 6), we define all key concepts that are extensively used in all five chapters.

# 6 A Framework for Modeling Semantic Similarity Based on Latent Cross-Lingual Topics

*Ever tried. Ever failed. No matter. Try again. Fail again. Fail better.*

— Samuel Beckett

## 6.1 Introduction

*Cross-lingual semantic word similarity* is the task of detecting words (or more generally, text units) that address similar semantic concepts and convey similar meanings across languages. Models of cross-lingual similarity are typically used to automatically induce *bilingual lexicons* and have found numerous applications in information retrieval (IR), statistical machine translation (SMT) and other natural language processing (NLP) tasks. Within the IR framework, the output of the cross-lingual models of semantic similarity is a key resource in the models of dictionary-based cross-lingual information retrieval [15, 41, 127, 107, 162, 124, 172, 319, 149, 296] or may be utilized in query expansion in cross-lingual IR models [1, 313]. These models of cross-lingual semantic similarity may also be utilized as an additional source of knowledge in SMT systems [295, 80, 229, 326, 190, 152]. Additionally, the models are a crucial component in the cross-lingual tasks involving a sort of cross-lingual knowledge transfer, where the knowledge about utterances in one language may be transferred to another. The utility of the transfer or annotation projection by means of

bilingual lexicons obtained from the models of cross-lingual semantic similarity has already been proven in various tasks such as semantic role labeling [232, 304], parsing [337, 83, 281], POS tagging [329, 63, 280, 97], inducing selectional preferences [234], named entity recognition [148], verb classification [198], named entity segmentation [97] and others.

In this part of the thesis, we investigate *distributional models of cross-lingual semantic similarity* from non-parallel data, where the key idea is to exploit an idea known as the *distributional hypothesis* [119] which states that words with similar meanings are likely to appear in similar *contexts*. In this chapter in particular, we show that latent cross-lingual topics provide a sound mathematical representation of this context and may be regarded as important context features. We measure semantic similarity between words $w_1^S$ given in $L_S$ and word $w_2^T$ in $L_T$ by the extent of how often they are present and how important they are over the same latent cross-lingual topics. In short, the word $w_2^T$ is semantically similar to $w_1^S$ if the distribution of $w_2^T$ over latent cross-lingual topics (extracted from per-topic word distributions for $L_T$) is similar to the probability distribution of $w_1^S$ over the topics (extracted from per-topic word distributions for $L_S$). In that respect, the contributions of this chapter are as follows:

(i) We demonstrate how to build a new latent cross-lingual semantic space spanned by latent cross-lingual topics induced from non-parallel data.

(ii) We propose, evaluate and compare new models of cross-lingual semantic similarity which rely on the knowledge of latent cross-lingual topics.

(iii) We propose and describe *topic space pruning*. By using only a subset of the most important cross-lingual topics (semantically most relevant features) in similarity calculations, we are able to significantly improve performance of our models of similarity in both accuracy and speed.

The remainder of the chapter is structured as follows. In sect. 6.2, we provide an overview of distributional models of cross-lingual similarity from comparable corpora; we present all key definitions, terminology and discuss shared cross-lingual features and related work. In sect. 6.3, we present our framework for modeling cross-lingual semantic similarity based on latent cross-lingual topics and describe new MuPTM-based models of similarity. Following that, sect. 6.4 discusses bilingual lexicon extraction as our evaluation task, and our experimental setup. Results and comparison of the models on the BLE task are provided in sect. 6.5. Finally, the main conclusions of this chapter are summarized in sect. 6.6.

## 6.2 Cross-Lingual Semantic Similarity: An Overview of Distributional Models

### 6.2.1 Definitions

Distributional approaches to detecting cross-lingual semantic word similarity from non-parallel data are all based on an idea known as the *distributional hypothesis* [119], which states that words with similar meanings are likely to appear in similar contexts. Each word is typically represented by a high-dimensional vector called *context vector* in a feature vector space or a so-called *semantic space*, where the dimensions of the vector are its *context features*. The semantic similarity of two words, $w_1^S$ given in the source language $L_S$ with vocabulary $V^S$ and $w_2^T$ in the target language $L_T$ with vocabulary $V^T$ is then:

$$sim(w_1^S, w_2^T) = SF(vec(w_1^S), vec(w_2^T)) \tag{6.1}$$

where $vec(w_1^S)$ is an $N$-dimensional context vector with $N$ context features $c_n$:

$$vec(w_1^S) = [sc_1^S(c_1), \ldots, sc_1^S(c_n), \ldots, sc_1^S(c_N)] \tag{6.2}$$

$sc_1^S(c_n)$ denotes a co-occurrence score for $w_1^S$ associated with context feature $c_n$ (similar for $w_2^T$). $SF$ is a similarity function operating on the context vectors.[1]

**Modeling Co-Occurrence: Weighting Functions.** As mentioned, given a word $w_1^S \in V^S$, $sc_1^S(c_n)$ assigns a co-occurrence score of the word $w_1^S$ with some context feature $c_n$. Distributional models differ in the way each $c_n$ is weighted, that is, the way the co-occurrence of $w_1^S$ and its context feature $c_n$ is mapped to the score $sc_1^S(c_n)$. There exists a variety of options for weighting: the values of $sc_1^S(c_n)$ are typically raw co-occurrence counts $C(w_1^S, c_n)$, conditional feature probability scores $P(c_n|w_1^S)$, weighting heuristics such as term frequency-inverse document frequency (TF-IDF), point-wise mutual information (PMI), or association scores based on hypothesis testing such as log-likelihood ratio (LLR).

**Obtaining Similarity Scores: Similarity Functions.** Once two words are represented as $N$-dimensional vectors in the same feature space (see eq. (6.2)), it is possible to measure their similarity in that feature space by means of a *similarity function*. There is a plethora of different similarity functions organized

[1] The reader has to be aware that the presentation of work in part III tackles the more difficult cross-lingual setting. We present a unified general probabilistic framework which does not change its modeling premises regardless of the actual setting (monolingual vs. cross-lingual or multilingual). All proposed models are fully functional in the monolingual setting. Monolingual models of similarity are special cases of cross-lingual models and are subsumed by this framework.

in different families according to [43]: (1) the inner product family of SF-s such as the cosine similarity used in [93, 160] or the Jaccard index [243, 123], (2) the Minkowski family, with SF-s such as the Euclidean distance or the city-block metric as used in [247], (3) the fidelity family, with SF-s such as the Bhattacharyya coefficient [144], the Shannon's entropy family, with SF-s such as the Kullback-Leibler divergence [312] or the Jensen-Shannon divergence [236], (4) the graph-based family, with SF-s such as SimRank [164], or (5) the family of SF-s tailored specifically for measuring semantic similarity such as the Lin Measure [177], etc. For an overview of these similarity functions and even more options, we refer the interested reader to the survey papers [167, 43].

**Output of Models of Semantic Similarity: Ranked Lists.** After applying a similarity function, for each source word $w_1^S$, we can build a *ranked list* $RL(w_1^S)$. The ranked list consists of all words $w_j^T \in V^T$ ranked according to their respective similarity scores $sim(w_1^S, w_j^T)$. In the similar fashion, we can build a ranked list $RL(w_2^T)$, for each target word $w_2^T$. We call the top $M$ best scoring target words $w_j^T$ for some source word $w_1^S$ its $M$ *nearest neighbors*. The ranked list for $w_1^S$ comprising only its $M$ nearest neighbors is called *pruned ranked list* (i.e., the ranked list is effectively pruned at position $M$), and we denote it as $RL_M(w_1^S)$. The single nearest cross-lingual neighbor for $w_1^S$ is called its *translation candidate*, and in case that is a word $w_2^T$, we write $TC(w_1^S) = w_2^T$.

One may construct a *one-to-one bilingual lexicon* from the output ranked lists of semantically similar words by simply harvesting all translation candidates, that is, by retaining all cross-lingual pairs $(w_1^S, TC(w_1^S))$. The pair $(w_1^S, TC(w_1^S))$ is referred to as a *translation pair* or a *bilingual lexicon entry*.

### 6.2.2 Related Work (Shared Cross-Lingual Features)

In order to compute cross-lingual semantic word similarity, one needs to design the context features of words given in two different languages that span a *shared cross-lingual semantic space* or a *shared cross-lingual vector space*. It means that words need to have the same representations over the same set of features irrespective of their actual language. Context vectors $vec(w_1^S)$ and $vec(w_2^T)$ for both source and target words are then compared in the shared semantic space independently of their respective languages. Such cross-lingual semantic spaces are typically spanned by:

(1) *Entries from an external bilingual lexicon* which is hand-crafted or extracted from a parallel corpus [246, 93, 247, 73, 92, 102, 213, 99, 160, 269, 6, 173, 174, 283]. These approaches presuppose existence of an expensive external resource in the form of a bilingual lexicon or parallel data, which is a rather heavy

assumption for many language pairs and domains for which such high-quality resources do not exist.

(2) *Predefined explicit cross-lingual categories* obtained from a knowledge base or an ontology [69, 94, 49, 121, 2, 122, 193]. The typical features are Wikipedia categories, Wikipedia anchors or categories from EuroWordNet [310]. A problem with these approaches again lies in the fact that it is extremely time-consuming and expensive to build such knowledge bases and ontologies for different languages, that is, they again presuppose existence of high-quality external resources which effectively limits their portability to other language pairs and domains. Moreover, it is especially challenging to realize such explicit structures cross-lingually and define shared cross-lingual categories.

(3) *Latent language-independent semantic concepts/axes* (e.g., latent cross-lingual topics) induced by an algebraic model [81, 159], or more recently by a generative probabilistic model [117, 64, 312]. These approaches are fully data-driven as they utilize only *internal evidence* from a given corpus. However, all previous approaches still rely on language pair specific knowledge such as orthographic clues [155, 117, 33, 64] or again require an initial bilingual lexicon [117, 33] in modeling.

In this part of the thesis we are interested in the models of similarity from item (3). In other words, we are interested in a *specific type of context features*, that is, latent cross-lingual semantic topics/concepts. In summary, *we explore the models of cross-lingual semantic similarity and build a new statistical framework in a particularly difficult (but extremely cheap) minimalist setting which builds only on co-occurrence counts and latent cross-lingual semantic topics/concepts induced directly from comparable corpora, and which does not rely on any other resource (e.g., machine-readable dictionaries, parallel corpora, explicit ontology and category knowledge).* In chapter 9, we also tackle the models from item (1), but contrary to the prior work, we will demonstrate how to build these feature sets without any parallel data or external bilingual lexicons to obtain shared cross-lingual features. In that chapter, we will bootstrap these shared features from an initial *seed set* of features obtained by an initial model of similarity built within our minimalist setting (and effectively remaining within the same cheap minimalist setting).

### 6.2.3 Quick Notes on Terminology

**Of Names and Naming Conventions.** The term *distributional models of semantic similarity* which we predominantly use in this thesis is *per se* rather vague. However, the reader must be aware that the relevant literature lists other terms that essentially refer to the exact same concept, such as *distributional semantic models*, *vector space models*, *semantic space models*, *word space models*.

**Of Similarity and Relatedness.** Even the term *semantic similarity* is vague as it may in general denote similarities between documents, words/phrases or relations [299]. This thesis tackles the problem of *attributional similarity* of words [298], which comprises standard taxonomic semantic relations such as synonymy (a relation between words with the same or similar meanings, e.g., *buy* and *purchase*), hyponymy and hypernymy (a hyponym is a word in a *type-of* relation with its hypernym, e.g., *pigeon* is a hyponym of *bird* which is in turn a hyponym of *animal*), co-hyponymy (e.g., *seagull* and *crow* are co-hyponyms of the shared hypernym *bird*), etc. [17]. Words like *cat* and *kitten*, for instance, are attributionally similar in the sense that their meanings share a large number of attributes: they are animals, they meow, they like to drink milk, etc. The here investigated attributional similarity is opposed to *relational similarity* which refers to detecting properties and relations shared between pairs of words, e.g., *cat-animal* and *car-vehicle*. Moreover, the reader has to be aware that the concept of semantic similarity is more specific than *semantic relatedness* (although the two are sometimes used interchangeably) as relatedness includes concepts such as antonymy (a relation between two words with completely opposite meanings, e.g., *war-peace*) and meronymy (a relation where one word is a constituent part of another, e.g., *finger-hand*), while similarity does not [39].

**Of Types and Tokens.** A *token* is a single instance of a word symbol, whereas a *type* is a general class of tokens [186]. If we take a quote from Samuel Beckett at the beginning of this chapter as an example, we say that word types such *Ever* or *again* occur once in the quote, while there are two tokens/instances of these word types occurring in the quote. A difference between type-based and token-based models of similarity is especially important for *polysemous* words (i.e., words that exhibit more than meaning such as *plant*, *shed*, *bank* or *match*), and we will extensively refer to that difference when building our context-sensitive models of cross-lingual similarity (chapter 10).

## 6.3 Cross-Lingual Semantic Similarity via Latent Cross-Lingual Topics

In this section, we show how to use latent cross-lingual topics to build models of cross-lingual semantic similarity. First, we present a set of models which measure the similarity according to the similarity of words' conditional topic distributions. Other models presented in this section explore other possibilities for exploiting the shared set $\mathcal{Z}$ of latent cross-lingual topics when constructing similarity models.

### 6.3.1 Conditional Topic Distributions

After training, a multilingual topic model outputs per-topic word distributions with probability scores $P_S(w_1^S|z_k)$ and $P_T(w_2^T|z_k)$, for each $w_1^S \in V^S$, $w_2^T \in V^T$ and $z_k \in \mathcal{Z}$. It holds that $\sum_{k=1}^{K} P_S(w_1^S|z_k) = 1$ and $\sum_{k=1}^{K} P_T(w_2^T|z_k) = 1$ [2], since each language has its own language-specific distribution over vocabulary words (see sect. 4.2.1 and sect. 4.4.3).

In order to quantify the similarity between two words $w_1^S \in V^S$ and $w_2^T \in V^T$, we may employ the same trick that has been used for obtaining the degree of similarity between two documents in the monolingual setting [278] and the cross-lingual setting [222]. Since each document $d_j$ is represented as a mixture of topics by means of per-document topic distributions given by the probability scores $P(z_k|d_j)$, the similarity between two documents can be established by measuring the similarity of these probability distributions. When dealing with the similarity of words $w_1^S$ and $w_2^T$, we need to measure the similarity of their respective *conditional topic distributions*, given by the probability scores $P(z_k|w_1^S)$ and $P(z_k|w_2^T)$, for each $z_k \in \mathcal{Z}$. Each word, regardless of its actual language, is then represented as a point in a $K$-dimensional latent semantic space. In other words, each word, irrespective to the language, is represented as a distribution over the $K$ latent topics/concepts, where the $K$-dimensional vector representation of $w_1^S \in V^S$ (similar for $w_2^T \in V^T$) is:

$$vec(w_1^S) = [P(z_1|w_1^S), \ldots, P(z_k|w_1^S), \ldots, P(z_K|w_1^S)] \tag{6.3}$$

Using Bayes' rule, we can compute these probability scores:

$$P(z_k|w_1^S) = \frac{P(w_1^S|z_k)P(z_k)}{P(w_1^S)} = \frac{P(w_1^S|z_k)P(z_k)}{\sum_{l=1}^{K} P(w_1^S|z_l)P(z_l)} \tag{6.4}$$

where $P(w_1^S|z_k)$ is known directly from the per-topic word distributions. $P(z_k)$ is the prior topic distribution which can be used to assign higher a priori importance to some cross-lingual topics from the set $\mathcal{Z}$ [136]. However, in a typical setting where we do not posses any prior knowledge about the corpus and the likelihood of finding specific latent topics in that corpus, we assume the uniform prior over latent cross-lingual concepts/topics [114] (i.e., that all topics/concepts are equally likely before we observe any training data). The probability scores $P(z_k|w_1^S)$ from eq. (6.4) for conditional topic distributions in

[2] A remark on notation throughout the rest of the thesis: Since the shared space of cross-lingual topics allows us to construct a uniform representation for all words regardless of their actual language, due to simplicity, we use notation $P(w_i|z_k)$ and later $P(z_k|w_i)$ instead of $P_S(w_i|z_k)$ or $P_S(z_k|w_i)$ (similar for subscript $T$). However, the reader must be aware that, for instance, $P(w_i|z_k)$ actually means $P_S(w_i|z_k)$ if $w_i \in V^S$, and $P_T(w_i|z_k)$ if $w_i \in V^T$.

that case may be further simplified:

$$P(z_k|w_1^S) = \frac{P(w_1^S|z_k)}{\sum_{l=1}^{K} P(w_1^S|z_l)} = \frac{\phi_{k,1}^S}{\sum_{l=1}^{K} \phi_{l,1}^S} = \frac{\phi_{k,1}^S}{Norm_{\phi_{:,1}^S}} \tag{6.5}$$

where we denote the normalization factor $\sum_{l=1}^{K} \phi_{l,1}^S$ as $Norm_{\phi_{:,1}^S}$. A similar derivation follows for each $w_2^T \in V^T$ and the similarity between two words may then be computed as the similarity between their conditional topic distributions as given by eq. (6.4) or eq. (6.5). We will use this property extensively in our models of cross-lingual similarity.

### 6.3.2 KL Model and JS Model

Now, once the conditional topic distributions are computed, any similarity metric may be used as SF to quantify the degree of similarity between the representations of words by means of these conditional topic distributions. We present and evaluate a series of models which employ the most popular SF-s reported in the relevant literature. Each SF in fact gives rise to a new model of cross-lingual semantic similarity!

The first model relies on the Kullback-Leibler (KL) divergence which is a common measure of (dis)similarity between two probability distributions [178]. The KL divergence of conditional topic distributions for two words $w_1^S$ and $w_2^T$ is an asymmetric measure computed as follows (our *KL model*):

$$sim(w_1^S, w_2^T) = KL(vec(w_1^S), vec(w_2^T)) = \sum_{k=1}^{K} P(z_k|w_1^S) \log \frac{P(z_k|w_1^S)}{P(z_k|w_2^T)} \tag{6.6}$$

$$= \sum_{k=1}^{K} \frac{\phi_{k,1}^S}{Norm_{\phi_{:,1}^S}} \log \frac{\phi_{k,1}^S \cdot Norm_{\psi_{:,2}^T}}{\psi_{k,2}^T \cdot Norm_{\phi_{:,1}^S}} \tag{6.7}$$

The Jensen-Shannon (JS) divergence [178, 59, 58] is a symmetric (dis)similarity measure closely related to the KL divergence, defined as the average of the KL divergence of each of two distributions to their average distribution. The similarity of conditional topic distributions is computed as follows (our *JS model*):

$$sim(w_1^S, w_2^T) = JS(vec(w_1^S), vec(w_2^T))$$

$$= \frac{1}{2}\Big( \sum_{k=1}^{K} P(z_k|w_1^S) \log \frac{P(z_k|w_1^S)}{P_{avg}(z_k|w_1^S)} + \sum_{k=1}^{K} P(z_k|w_2^T) \log \frac{P(z_k|w_2^T)}{P_{avg}(z_k|w_2^T)} \Big) \tag{6.8}$$

where $P_{avg}(z_k|w_1^S) = P_{avg}(z_k|w_2^T) = \frac{P(z_k|w_1^S)+P(z_k|w_2^T)}{2}$, for each $z_k \in \mathcal{Z}$. Both KL and JS output non-negative scores, where a lower output score implies a lower divergence between two words, and therefore, a closer semantic similarity. Both KL and JS are defined only if they deal with real probability distributions, that is, if probability scores sum up to 1, Additionally, $P(z_k|w_1) > 0$ and $P(z_k|w_2) > 0$ has to hold for each $z_k$. Conditional topic distributions satisfy all these conditions.

A ranked list $RL(w_1^S)$ may be obtained by sorting words $w_2^T \in V^T$ in ascending order based on their respective dissimilarity/divergence scores computed by eq. (6.7) (KL) or eq. (6.8) (JS).

### 6.3.3 TCos Model

The next distance measure is the cosine similarity, which is one of the most popular choices for SF in distributional semantics [93, 40, 299]. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine similarity of conditional topic distributions (our *TCos model*) is computed as follows:

$$\begin{aligned} sim(w_1^S, w_2^T) &= TCos(vec(w_1^S), vec(w_2^T)) \\ &= \frac{\sum_{k=1}^{K} P(z_k|w_1^S)P(z_k|w_2^T)}{\sqrt{\sum_{k=1}^{K} P(z_k|w_1^S)^2} \cdot \sqrt{\sum_{k=1}^{K} P(z_k|w_2^T)^2}} \end{aligned} \tag{6.9}$$

The higher the score in the range $[0, 1]$ (all dimensions of our vectors are positive numbers and therefore the lower similarity bound is 0 instead of $-1$), the higher the similarity between two words. A ranked list $RL(w_1^S)$ may be obtained by sorting words $w_2^T \in V^T$ in descending order based on their respective similarity scores computed by eq. (6.9).

### 6.3.4 BC Model

Another similarity measure is the Bhattacharyya coefficient (BC) [23, 144]. The similarity of two words based on this similarity measure is defined as follows (our *BC model*):

$$sim(w_1^S, w_2^T) = BC(vec(w_1^S), vec(w_2^T)) = \sum_{k=1}^{K} \sqrt{P(z_k|w_1^S)P(z_k|w_2^T)} \tag{6.10}$$

In general, it measures the amount of overlap between two statistical samples which, unlike for KL and JS, do not have to be described by proper probability distributions. A higher score again implies a stronger semantic similarity between two words. The utility of the BC measure has not been investigated well in the literature on distributional semantics. Our experiments will reveal its potential in identifying semantically similar words across languages.

### 6.3.5 Cue Model

Another way of utilizing per-topic word distributions is to directly model the probability $P(w_2^T|w_1^S)$, where semantically most similar target words should have the highest probability to be generated as a response to a cue source word. The probability $P(w_2^T|w_1^S)$ emphasizes the (cross-lingual) associative relation between words [114]. Again, under the assumption of uniform topic prior, we can decompose the probability $P(w_2^T|w_1^S)$ as follows (our *Cue model*):

$$sim(w_1^S, w_2^T) = Cue(vec(w_1^S), vec(w_2^T)) = \sum_{k=1}^{K} P(w_2^T|z_k)P(z_k|w_1^S) \tag{6.11}$$

The probability value directly provides the degree of semantic similarity and a ranked list $RL(w_1^S)$ may be obtained by sorting words $w_2^T \in V^T$ in descending order based on their respective probability scores computed by eq. (6.11). The asymmetric Cue model has a strong theoretical foundation in cognitive science as it closely resembles the actual process in the human brain [219, 279, 114]. We will discuss the model and its properties in more detail in chapter 8, as it will serve as the base model for more advanced and robust models of similarity discussed there.

### 6.3.6 TI Model

The next model moves away from utilizing conditional topic distributions explicitly and aims to exploit latent cross-lingual topics in a different way. It builds a context vector $vec(w_1^S)$ as follows:

$$vec(w_1^S) = [\text{TTF-ITF}(w_1^S, z_1), \ldots, \text{TTF-ITF}(w_1^S, z_K)] \tag{6.12}$$

TTF-ITF (*term-topic frequency - inverse topic frequency*) is a novel weighting scheme which is analogous to and directly inspired by the TF-IDF (*term frequency - inverse document frequency*) weighting scheme in IR [186, 185]. Instead of conditional topic probability scores $P(z_k|w_1^S)$ the context features $sc_1(c_k)$ are now TTF-ITF scores. In our TTF-ITF weighting scheme, the

$\text{TTF}(w_1^S, z_k)$ part of the complete score $\text{TTF-ITF}(w_1^S, z_k)$ measures importance of $w_1^S$ for the particular topic $z_k$. It denotes the number of assignments of the latent cross-lingual topic $z_k$ to the occurrences of $w_1^S$ in the whole training corpus (i.e., that number is exactly the Gibbs count variable $v^S_{k,w_1^S}$ which is one of the variables utilized to obtain the output per-topic word distributions in sect. 4.4.3). The $\text{ITF}(w_1^S)$ score measures global importance of $w_1^S$ across all latent cross-lingual topics. Words that are prominent for only a small subset of topics from $\mathcal{Z}$ are given higher importance for these topics as such words are in general more descriptive for these specific topics than high-frequency words that occur frequently over all topics. The inverse topic frequency for the word $w_1^S$ across the set of cross-lingual topics is computed as:[3]

$$ITF(w_1^S) = \log \frac{K}{1 + |\{z_k : v^S_{k,w_1^S} > 0\}|} \tag{6.13}$$

The final $\text{TTF-ITF}(w_1^S, z_k)$ score for the source language word $w_1^S$ and the topic $z_k$ is then calculated as $\text{TTF-ITF}(w_1^S, z_k) = \text{TTF}(w_1^S, z_k) \cdot \text{ITF}(w_1^S)$. Once the same vector representation for $w_2^T \in V^T$ has been obtained, the similarity between words $w_1^S$ and $w_2^T$ may again be computed by means of their $K$-dimensional vector representations from eq. (6.12) using the cosine similarity as follows (our *TI* model):

$$\begin{aligned} sim(w_1^S, w_2^T) &= TI(vec(w_1^S), vec(w_2^T)) \\ &= \frac{\sum_{k=1}^{K} \text{TTF-ITF}(w_1^S, z_k) \cdot \text{TTF-ITF}(w_2^T, z_k)}{\sqrt{\sum_{k=1}^{K} \text{TTF-ITF}(w_1^S, z_k)^2} \cdot \sqrt{\sum_{k=1}^{K} \text{TTF-ITF}(w_1^S, z_k)^2}} \end{aligned} \tag{6.14}$$

Note that, since this model directly utilizes the word-topic matrix, other weighting schemes besides TTF-ITF such as standard PMI or LLR may be utilized to build feature scores in the word vectors.

### 6.3.7 TI+Cue Model

In the original paper [312] which is the basis of this chapter, we have discussed that the Cue model and the TI model interpret and exploit the shared set of latent cross-lingual topics in different ways. Therefore, by combining the two models and capturing different evidences of similarity, we should be able to boost the quality of obtained ranked lists. As in [312], we present a linear

[3] A stronger association with a latent cross-lingual topic is modeled by setting a higher *threshold* value in $v^S_{k,w_i^S} > threshold$, where we have chosen a threshold value 0.

combination of the two models (with $\gamma$ as the interpolation parameter), where the overall score is computed as follows (our *TI+Cue model*):

$$sim(w_1^S, w_2^T) = \gamma sim_{TI}(w_1^S, w_2^T) + (1-\gamma) sim_{Cue}(w_1^S, w_2^T) \tag{6.15}$$

### 6.3.8 Topic Space Pruning

All these models of similarity have a straightforward theoretical explanation - they assign high similarity scores for pairs of words that assign similar importance to the same latent cross-lingual topics/concepts, that is, the same axes in the shared semantic space. In the core of all models of similarity are point-wise additive formulas, i.e., the models perform calculations over each cross-lingual topic $z_k \in \mathcal{Z}$ and each calculation contributes to the overall sum. However, each word is usually important for only a limited number of topics/concepts. For models that make use of conditional topic distributions it means that words exhibit high conditional topic probability scores for only a small subset of cross-lingual topics. In a typical setting for mining semantically similar words using latent topic models in both monolingual [114, 77] and cross-lingual settings [312], the best results are obtained with the number of topics set to a few thousands ($\approx 2000$, more analysis follows in the upcoming sections).

For the sake of simplicity, the presentation in this section is valid for models that utilize conditional topic distributions with scores $P(z_k|w_1^S)$. Since $P(z_k|w_1^S) > 0$ for each $z_k \in \mathcal{Z}$, a lot of probability mass is assigned to latent topics that are not relevant to the given word. Reducing the dimensionality of the semantic representation a posteriori to only a smaller number of the most informative semantic axes in the latent space should decrease the effects of that statistical noise, and even more firmly emphasize the latent correlation among words. The utility of such *semantic space pruning* in monolingual settings [250] was also detected previously for latent semantic models [159, 114, 314].

Given two words $w_1^S$ and $w_2^T$, we prune the representation of these words in the shared latent semantic space spanned by cross-lingual topics as summarized in alg. 6.1. Both $w_1^S$ and $w_2^T$ are now represented by their $K'$-dimensional context vectors: for $w_1^S$ the vector is $vec(w_1^S) = [P(z'_1|w_1^S), \ldots, P(z'_{K'}|w_1^S)]$, where context features are now the semantically most relevant cross-lingual topics for the word $w_1^S$. We may again employ any of SF-s (e.g., KL, JS, BC, TCos) on these reduced representations, that is, pruned context vectors with the adjusted conditional topic probability scores to calculate similarity.

**Algorithm 6.1**: TOPIC SPACE PRUNING

1: **obtain** a subset $\mathcal{Z}_{K'} \subseteq \mathcal{Z}$ of $K' \leq K$ latent cross-lingual topics with the highest values $P(z_k|w_1^S)$. Calculating the similarity score $sim(w_1^S, w_2^T)$ may be interpreted as: "Given a word $w_1^S$ detect how similar another word $w_2^T$ is to the word $w_1^S$." Therefore, when calculating $sim(w_1^S, w_2^T)$, even when dealing with symmetric similarity functions such as JS or BC, we always consider only the ranking of scores $P(z_k|w_1^S)$ for pruning. The subset $\mathcal{Z}_{K'}$ is then $\{z'_1, \ldots, z'_{K'}\}$ ;
2: **retain** conditional topic probability scores $P(z'_k|w_1^S)$ for the word $w_1^S$ only over topics $z'_k \in \mathcal{Z}_{K'}$ ;
3: **retain** conditional topic probability scores $P(z'_k|w_2^T)$ for $w_2^T$ over the same cross-lingual topics $z'_k \in \mathcal{Z}_{K'}$ ;

## 6.4 Experimental Setup

### 6.4.1 Evaluation Task: Bilingual Lexicon Extraction

All context-insensitive models of cross-lingual semantic similarity in this chapter and following chapters are evaluated in the task of *bilingual lexicon extraction (BLE)*, which is the standard evaluation task in the cross-lingual setting and is extensively used as the main evaluation task in the related literature dealing with comparable corpora (e.g., [93, 247, 155, 102, 134, 6, 174, 235, 283] and many more).

The goal of the task is to build a *non-probabilistic bilingual lexicon of one-to-one word translations*. We may obtain this lexicon by harvesting only one-to-one translation pairs $(w_1^S, TC(w_1^S))$. For evaluation purposes, we also retain pruned ranked lists $RL_M(w_1^S)$ at position $M$, and evaluate how high (or low) in these lists we observe the actual top translation candidate. That score will also provide a valuable insight into the quality of our models of cross-lingual semantic similarity.

### 6.4.2 Training, Testing and Evaluation

**Training Data.** In this chapter, we evaluate and compare models of similarity in the Italian-English BLE task. We train on the collection of 18,898 Italian-English Wikipedia article pairs (see sect. 3.2.2 in chapter 3). Following prior work (e.g., [155, 117, 243]), we use TreeTagger [267] for POS-tagging and lemmatization of the corpora, and then retain only nouns that occur at least 5 times in the corpus. We record the lemmatized form when available, and the original form otherwise. Our final vocabularies consist of 7,160 Italian nouns and 9,116 English nouns.

**Multilingual Topic Model.** We have trained the BiLDA model on the IT-EN Wikipedia data with standard recommended hyper-parameter settings: $\alpha = 50/K, \beta = 0.01$ [278]. We have trained the BiLDA model using the Gibbs sampling training procedure (see sect. 4.4.2 in chapter 4). We have varied the number of topics $K$ for BiLDA from 200 to 3,500 with steps of 300 or 400 to measure the influence of the parameter $K$ on the overall scores, that is, to test how the granularity of the shared topical space influences the quality of our models of similarity.

**Test Data and Ground Truth.** Since our task is bilingual lexicon extraction, we designed a set of ground truth one-to-one translation pairs $G$ similarly to the procedure already described in sect. 5.4 in chapter 5. We randomly sampled a set of 1,000 Italian nouns from our Wikipedia corpora (i.e., we conduct our experiments in the Italian-to-English direction) which are to be regarded as our *test words*. Following that, we used the *Google Translate* tool plus an additional annotator to translate those words to English. The annotator manually revised the lists and retained only words that have their corresponding translation in the English vocabulary. Additionally, only one possible translation was annotated as correct. When more than one translation is possible (e.g., when dealing with polysemous words), the annotator marked as correct the translation that occurs more frequently in the English part of our Wikipedia training data. Following the same procedure, we also sampled 200 Italian words which are not in the test set and obtained their one-to-one word translations. These 200 one-to-one translation pairs constitute our *development set* which we utilize in this and all subsequent chapters to tune the parameters of our models in all further experiments. In this chapter, we only have to tune the interpolation parameter $\gamma$ in the TI+Cue model which is set to 0.1.

**Evaluation Metrics.** We measure the performance on the BLE task using a standard *Top M* accuracy ($Acc_M$) metric [102, 283]. It denotes the number of source words $w_i^S$ from ground truth translation pairs $(w_i^S, GTC(w_i^S))$ whose list $RL_M(w_1^S)$ contains the correct translation according to our ground truth over the total number of ground truth translation pairs ($|G| = 1000$). $GTC(w_i^S)$ denotes the translation candidate for $w_i^S$ as given in the ground truth. More formally, $Acc_M$ is computed as follows:

$$Acc_M = \frac{1}{|G|}|\{w_i^S : w_i^S \in G \wedge GTC(w_i^S) \in RL_M(w_i^S)\}| \tag{6.16}$$

Since we can build a one-to-one bilingual lexicon by simply harvesting one-to-one translation pairs $(w_1^S, TC(w_1^S))$, the quality of the actual lexicon is best reflected in the $Acc_1$ score. In some experiments, we also report the *mean reciprocal rank (MRR)* [309], a standard NLP/IR metric (also reported in chapter 5) which, unlike $Acc_M$, assigns larger weights if a correct translation candidate is found higher in the ranked list.

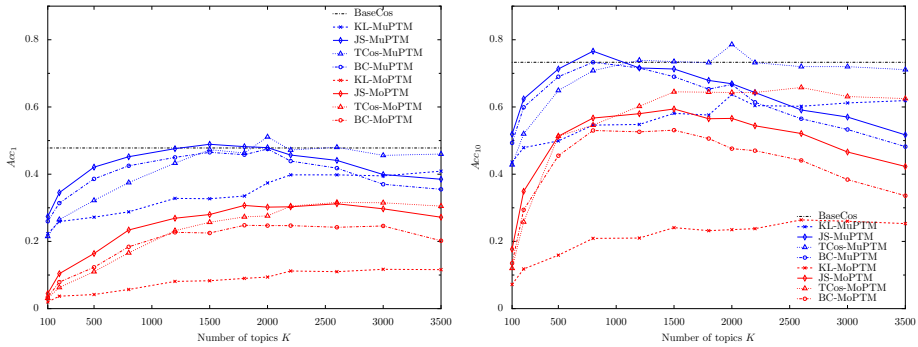
**Models for Comparison.** We evaluate all our MuPTM-based models of similarity in the BLE task (their codes are **-MuPTM*, e.g., KL-MuPTM or JS-MuPTM). We compare them against baseline models which also exploit document alignments when mining semantically similar words and translation candidates from comparable corpora:

(1) The first set of models utilizes standard monolingual LDA [31] (see sect. 4.4) on concatenated aligned Wikipedia articles. The LDA models are trained with exactly the same parameter setup as our BiLDA models. By the concatenation of aligned Wikipedia articles given in the source and the target language, we effectively remove the gap between the languages and train on the obtained set of "merged" documents and acquire a shared set of latent topics represented by words in both languages. However, we may still distinguish between source language words and target language words, and use only a subset of all words comprising all target language words in final ranked lists. The goal of this comparison is to test whether it is useful to provide separate topic representations in two languages by jointly training on separate documents (see sect. 4.6). We again test all models from the previous sections as with MuPTM. Since the obtained set of models effectively relies on a monolingual topic model, their codes are **-MoPTM*, e.g., KL-MoPTM, JS-MoPTM.

(2) Another baseline model is conceptually similar to our TI model. The baseline model computes word vectors, but now in the original word-document space (instead of the lower-dimensional word-topic space) using the TF-IDF weighting scheme (which is completely analogous to the TTF-ITF weighting scheme, see sect. 6.3.6) and the cosine similarity on the obtained word vectors. This comparison serves to test whether we gain some extra contextual information by translating our problem from the word-document to the word-topic space, besides the obvious fact that we produce a sort of *dimensionality reduction* which in turn speeds up computations. In other words, we test whether topic models have the ability to build clusters of words which might not always co-occur together in the same textual units and therefore add extra information of similarity besides a direct co-occurrence captured by this baseline model (*BaseCos*).

## 6.5 Experiments, Results and Discussion

We conduct two different batches of experiments: (1) We compare all our proposed models against the baseline models from sect. 6.4 and measure the influence of the number of topics $K$ on the overall results in the BLE task; (2) We test and report the effect of topic space pruning for a selection of models.




(a) $Acc_1$ scores (b) $Acc_{10}$ scores

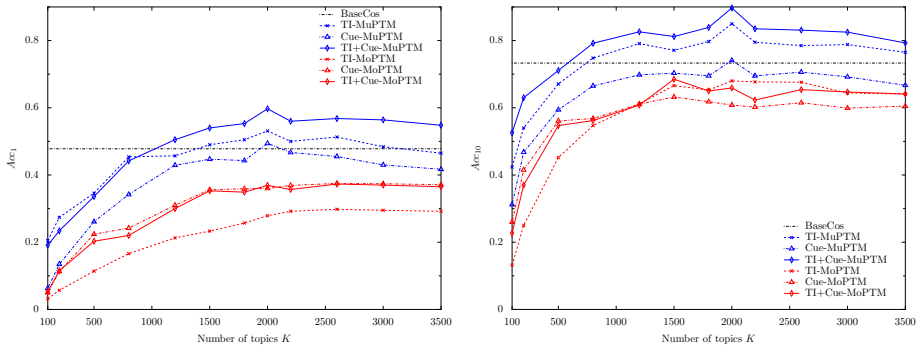
Figure 6.1: $Acc_1$ and $Acc_{10}$ scores for KL-MuPTM, JS-MuPTM, TCos-MuPTM, BC-MuPTM, and comparisons with baseline models: KL-MoPTM, JS-MoPTM, TCos-MoPTM, BC-MoPTM, and BaseCos.

### 6.5.1 Experiment I: Comparison of All Models

$Acc_1$ and $Acc_{10}$ scores in the BLE task for all models relying directly on the similarity function operating on context vectors with conditional topic probability scores (KL-*, JS-*, BC-*, TCos-* models) are displayed in fig. 6.1a and fig. 6.1b respectively. $Acc_1$ and $Acc_{10}$ for all other models are displayed in fig. 6.2a and fig. 6.2b respectively. Additionally, tab. 6.1 lists the best results for all models along with the optimal number of topics $K$ with which these results have been obtained. Based on all these results, we may derive a series of important conclusions:

(i) A comparison of all *-MuPTM models (blue lines) and all *-MoPTM models (red lines) clearly reveals the utility of training BiLDA on separate documents in place of training standard LDA on concatenated documents. All MuPTM-based models significantly outscore their MoPTM-based variants with LDA trained with exactly the same parameters as BiLDA. By training LDA on concatenated documents, we inherently introduce imbalance in the model, since one of the languages might clearly dominate the latent topic estimation (e.g., in cases when, for instance, English data is of higher quality than Italian data).

(ii) The choice of a similarity function matters. If we compare strictly SF-s operating with exactly the same representations (context vectors comprising conditional topic distributions) as given in fig. 6.1a and fig. 6.1b, we may observe that the KL model is significantly outperformed by the related JS model (which effectively performs a sort of symmetrization) and two other novel similarity models (the TCos model and the BC model) operating with exactly the same word representations.




(a) $Acc_1$ scores (b) $Acc_{10}$ scores

Figure 6.2: $Acc_1$ and $Acc_{10}$ scores for Cue-MuPTM, TI-MuPTM, TI+Cue-MuPTM, and comparisons with baseline models: Cue-MoPTM, TI-MoPTM, TI+Cue-MoPTM, and BaseCos.

| **Topic model:** | **-MuPTM** | | **-MoPTM** | |
|---|---|---|---|---|
| **Similarity model** | $Acc_1$ $(K)$ | $Acc_{10}$ $(K)$ | $Acc_1$ $(K)$ | $Acc_{10}$ $(K)$ |
| KL | 0.409 (3500) | 0.619 (3500) | 0.117 (3000) | 0.264 (2600) |
| JS | 0.489 (1500) | 0.766 (800) | 0.312 (2600) | 0.594 (1500) |
| TCos | 0.511 (2000) | 0.786 (2000) | 0.316 (2600) | 0.658 (2600) |
| BC | 0.475 (2000) | 0.733 (1200) | 0.248 (1800) | 0.531 (1500) |
| Cue | 0.494 (2000) | 0.741 (2000) | 0.375 (2600) | 0.615 (2600) |
| TI | 0.531 (2000) | 0.850 (2000) | 0.298 (2600) | 0.680 (2000) |
| TI+Cue | 0.597 (2000) | 0.897 (2000) | 0.373 (2600) | 0.685 (1500) |
| BaseCos | 0.478 (-) | 0.733 (-) | - | - |

Table 6.1: Best $Acc_1$ and $Acc_{10}$ scores over all values of $K$ (in parentheses after each result) for all compared models.

(iii) Based on these initial results, the TI model which relies on the representation with the new TTF-ITF weighting scheme is the best scoring basic model of similarity. However, we will later show that by pruning the topic space, other basic models such as JS and BC may outperform the TI model. Moreover, it is very interesting to note that the TI model, which is effectively the same model as BaseCos, but with a shift from the original word-document space to the newly induced word-topic space, outscores the BaseCos model. All other MuPTM-based models (except for KL-MuPTM) are at least on a par with BaseCos. This insight shows that reducing the dimensionality of the feature space in word representations and translating the problem from the original word-document space to this novel word-topic space might lead to both more effective and computationally faster models of similarity.

| (1) **romanzo** (novel) | (2) **paesaggio** (landscape) | (3) **cavallo** (horse) |
|---|---|---|
| writer | tourist | *horse* |
| novella | painting | stud |
| novelette | *landscape* | horseback |
| humorist | local | hoof |
| novelist | visitor | breed |

Table 6.2: Lists of the top 5 semantically similar words (Italian to English), where the correct translation candidate is not found (column 1), lies hidden lower in the pruned ranked list (2), and is retrieved as the most similar words (3). All three lists are obtained with TI+Cue-MuPTM.

(iv) We may observe that by combining the TI-MuPTM model and the Cue-MuPTM model, we are able to boost the overall performance. The results of the combined TI+Cue-MuPTM model outperform the results obtained by using any of the component models alone. The combined TI+Cue model displays the best overall performance across all compared models of similarity.

(v) Our models of similarity reach their optimal performances with larger values of $K$ (e.g., around the $2,000$ topics mark). While the tasks that required only coarse categorizations, such as event-based news clustering [67] or document classification [68] typically used a lower number of topics (in the [50-200] interval), cross-lingual semantic word similarity and bilingual lexicon extraction require a set of fine-grained latent cross-lingual topics which consequently leads to finer-grained topical representations of words. Based on these results, in all further experiments, we will set $K = 2000$, unless noted otherwise.

(vi) Additionally, it has been noted for both monolingual [299] and cross-lingual settings [235] that for distributional models synonymy is not the only semantic relation detected within the (pruned) ranked lists of words. The same is true for our distributional models relying on topical knowledge. For instance, besides direct cross-lingual synonymy, that is, the actual translational equivalence, we observe other semantic relations with words ranked highly in the lists (in top ten candidate words): near-synonymy (e.g., *incidente (accident) - crash*), antonymy (e.g., *guerra (war) - peace*), hyponymy (e.g., *particella (particle) - boson*), hypernymy (e.g., *ragazzo (boy) - child*), meronymy (e.g., *soldato (soldier) - troop*), holonymy (e.g., *mazzo (deck) - card* ) and other, uncategorized semantic relations (e.g., *vescovo (bishop) - episcopate*). The quantitative analysis (as performed in [235]) of the semantic relations detected by the models is beyond the scope of this work and will not be further investigated. Ranked lists of semantically similar words provide comprehensible and useful contextual information in the target language given a source word, even when the correct

translation candidate is missing, as might be seen in tab. 6.2. We will later exploit this finding when building information retrieval models (chapter 11).

### 6.5.2 Experiment II: Analysis of Topic Space Pruning

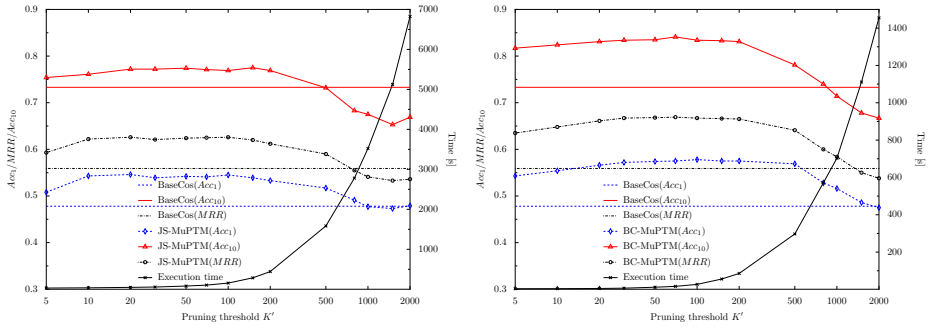
In the next set of experiments, we analyze the influence of topic space pruning on the behavior of our MuPTM-based models of cross-lingual similarity. All models use output per-topic word distributions from the BiLDA model trained with $K = 2000$ topics. Tab. 6.3 displays the results over different values for the pruning parameter $K'$. For the sake of clear presentation, we omit the results for: (1) the KL model whose behavior resembles the behavior of the JS model, only with lower overall scores, (2) the Cue model where we have not detected any major influence on the overall scores (i.e., the pruning is useful since it reduces execution time, but it does not lead to any improvements in scores), and (3) TI and TI+Cue models which rely on the cosine similarity and whose behavior resembles the behavior of the TCos model. Additionally, fig. 6.3a and fig. 6.3b display the change in overall scores for JS and BC over different values of $K'$ along with execution times for all pruned JS and BC models. These time-related experiments were conducted on a Intel(R) Xeon(R) CPU E5-2667 2.9GHz processor. We may notice several interesting phenomena:

(i) Topic space pruning helps to obtain higher results in the BLE task for the JS model (the increase is 7.5% even when probability scores do not sum up to 1, see sect. 6.3.2) and the BC model (the increase is 21.7%). Using only a small subset of possible features (e.g., $K' = 10$), we are able to acquire bilingual lexicons of higher quality with these models as reflected in $Acc_1$ scores. The improvement in $Acc_{10}$ and $MRR$ scores is also prominent. Moreover, as fig. 6.3a and fig. 6.3b reveal, the utility of topic space pruning is especially visible when we compare execution times needed to retrieve ranked lists for test words. For instance, while the BC model needs 1454.3 seconds to obtain ranked lists when we operate with full $K$-dimensional representations, the execution time is only 4.2 seconds with $K' = 10$, and we even obtain better results.

(ii) The reader might wonder why it is useful to induce a fine-grained latent topic space with a large number of topics, and then to perform pruning of the space afterwards, that is, to select only a small subset of the most informative features to represent words. The answer is as follows: While we desire to have a semantic space in which a large number of linguistic phenomena and topics are covered (a large set $\mathcal{Z}$), only a small subset of these topics is relevant to a particular word (topic space pruning). For instance, although we require that our semantic space is expressive enough to present topics and words related to *marine biology* or *radioactive isotopes*, these topics are completely irrelevant when we build a representation for a word *playmaker*.

| | **JS-MuPTM** | | | **TCos-MuPTM** | | | **BC-MuPTM** | | |
|---|---|---|---|---|---|---|---|---|---|
| $K'$ | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{20}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ |
| 1 | 0.380 | 0.488 | 0.704 | 0.0 | 0.0 | 0.0 | 0.477 | 0.575 | 0.760 |
| 2 | 0.454 | 0.543 | 0.717 | 0.113 | 0.142 | 0.195 | 0.497 | 0.601 | 0.792 |
| 5 | 0.508 | 0.593 | 0.754 | 0.191 | 0.232 | 0.301 | 0.543 | 0.635 | 0.817 |
| 10 | 0.543 | 0.622 | 0.761 | 0.210 | 0.243 | 0.303 | 0.554 | 0.648 | 0.824 |
| 20 | 0.546 | 0.626 | 0.772 | 0.207 | 0.245 | 0.314 | 0.566 | 0.661 | 0.831 |
| 30 | 0.539 | 0.621 | 0.772 | 0.220 | 0.260 | 0.329 | 0.572 | 0.667 | 0.834 |
| 50 | 0.542 | 0.624 | 0.774 | 0.267 | 0.318 | 0.427 | 0.574 | 0.668 | 0.835 |
| 70 | 0.541 | 0.625 | 0.771 | 0.315 | 0.375 | 0.502 | 0.575 | **0.669** | **0.844** |
| 100 | 0.545 | 0.626 | 0.769 | 0.357 | 0.425 | 0.572 | **0.578** | 0.667 | 0.834 |
| 150 | 0.539 | 0.620 | 0.775 | 0.394 | 0.471 | 0.639 | 0.575 | 0.666 | 0.833 |
| 200 | 0.533 | 0.612 | 0.769 | 0.408 | 0.490 | 0.670 | 0.575 | 0.665 | 0.831 |
| 500 | 0.517 | 0.590 | 0.732 | 0.471 | 0.559 | 0.739 | 0.569 | 0.641 | 0.781 |
| 800 | 0.491 | 0.555 | 0.683 | 0.483 | 0.573 | 0.748 | 0.528 | 0.600 | 0.740 |
| 1000 | 0.477 | 0.536 | 0.653 | 0.497 | 0.586 | 0.757 | 0.516 | 0.584 | 0.714 |
| 1500 | 0.463 | 0.573 | 0.692 | 0.512 | 0.598 | 0.769 | 0.486 | 0.550 | 0.678 |
| 2000 | 0.508 | 0.571 | 0.699 | 0.511 | 0.605 | 0.786 | 0.475 | 0.538 | 0.667 |

Table 6.3: Topic space pruning: $Acc_1$, $MRR$, and $Acc_{10}$ scores for JS-MuPTM, TCos-MuPTM and BC-MuPTM which rely on word representations by means of conditional topic distributions over different values of pruning parameter $K'$. BiLDA. $K = 2000$.




(a) JS model (b) BC model

Figure 6.3: $Acc_1$, $Acc_{10}$, $MRR$ scores for JS-MuPTM and BC-MuPTM along with their execution times. The horizontal axis is in logarithmic scale.

(iii) We have detected that topic space pruning for similarity models relying on the cosine similarity (TCos, TI, TI+Cue) negatively affects the performance. In the cosine similarity, since the normalization of an inner product is performed (unlike in BC), the absolute weights/probability scores are neglected and the

direction (instead of the absolute score) in the semantic space dominates the similarity function [139]. With a limited number of dimensions, the semantic space is not expressive enough and simply assigns high scores for many irrelevant target language words whose vectors are proportionally similar to the given source language word. For instance, take an extreme case where only one feature is left after pruning. The cosine similarity will assign a perfect similarity score of 1.0 for each target language word regardless of the actual absolute weights, while BC will not produce equal similarity scores for all words.

(iv) The best overall results are obtained with the BC model with $K' = 100$, and we may observe a major improvement over the baseline BaseCos model. A similar behavior is observed with the JS model. Moreover, the BC model is also the fastest model of all proposed models (e.g., 26.9 seconds compared to 51.5 seconds of TCos and 155.3 seconds of JS with $K' = 100$). In summary, by performing topic space pruning, we are able to improve our models of cross-lingual similarity both in terms of accuracy and speed. In all further experiments where we perform topic space pruning, unless noted otherwise, we will work with the $K' = 100$ best dimensions in the semantic space.

## 6.6 Conclusions and Future Work

In this chapter we have again tackled research question RQ2, but now in a setting without parallel data. To provide a solution to the question, we have presented a new framework for modeling cross-lingual semantic similarity which is supported by the knowledge of latent cross-lingual topics induced directly from non-parallel data. The usage of latent cross-lingual topics in modeling of cross-lingual semantic similarity is a completely new approach pioneered within this thesis. The proposed framework is purely statistical and fully corpus-based as it relies only on co-occurrence counts and latent cross-lingual topics which can be directly estimated from comparable data such as aligned Wikipedia articles given in different languages. The approach does not make any additional language-pair dependent assumptions, that is, it does not rely on a bilingual lexicon, orthographic clues or predefined ontology/category knowledge, and it does not require parallel data (recall requirement R4 and research question RQ2). That makes it portable to other language pairs (we will experiment with more language pairs in the upcoming chapters). Moreover, *while the focus of this chapter and thesis in general is on cross-lingual models similarity that we deem more general, all the models described in this chapter are fully operational in simpler monolingual settings.* The nature of presentation in this chapter (and all following chapters) regards monolingual models of semantic similarity within this framework only as special degenerate cases of the cross-lingual models which operate with only one language.

In this chapter, we have provided all key definitions and modeling assumptions (e.g., similarity functions, weighting functions, ranked lists, translation candidates), which will be used in the following chapters on cross-lingual semantic similarity. We have proposed, evaluated and described a series of models of cross-lingual similarity that aim to explore the set of latent cross-lingual topics in different ways. In the course of building this framework, we have made several important contributions: (1) We have shown that a transition from the original word-document space to the induced latent word-topic space yields more effective and computationally faster models of similarity; (2) We have proven the utility of training multilingual topic models on separate documents in two languages as much better results are obtained with BiLDA than with LDA trained on concatenated document pairs; (3) We have provided a systematic comparison of all our new models; (4) We have demonstrated the utility of topic space pruning.

The paths of future work are plentiful, and some of them will be followed in the upcoming chapters. For instance, extensions of this framework might include algorithms for selection of only highly reliable translation pairs (chapter 7) if the actual task is to deliver bilingual lexicons. The noisier but more abundant ranked lists of cross-lingual semantically similar words, such as the lists obtained by the models from this chapter, might prove more useful in another task, that is, in the cross-lingual information retrieval setting (see later in chapter 11). Other paths of future work also include building more robust models of similarity that rely on the basic models discussed in this chapter (chapter 8), using the knowledge from these models of similarity to obtain initial bilingual lexicons for bootstrapping approaches (chapter 9), or extending the framework towards context-sensitive models of similarity (chapter 10). Other possibilities include experimenting with other weighting schemes, similarity functions and other word representations within this framework, or designing models that are able to focus on a specific semantic relation (e.g., hyponymy and hypernymy). Moreover, since the proposed framework for modeling MuPTM-based cross-lingual semantic similarity is generic and completely topic model-independent, it allows experimentations with other multilingual topic models (see sect. 4.6 and sect. 4.7).

## 6.7 Related Publications

[1] **I. Vulić**, W. De Smet and M.-F. Moens. "Identifying word translations from comparable corpora using latent topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, Portland, Oregon, USA, 19-24 June 2011, pp. 479-484, ACL, 2011.

# 7 Selecting Highly Confident Translation Pairs

*Fast is fine, but accuracy is everything.*
— Xenophon (also attributed to Wyatt Earp)

## 7.1 Introduction

In the previous chapter, we have proposed new models of cross-lingual semantic similarity induced from non-parallel data. However, reported results on the task of bilingual lexicon extraction show that there is still ample room for improvement. Since comparable corpora construct a very noisy environment (e.g., some translation candidates are simply missing from the data, some words are low-frequency and are therefore difficult to capture by statistical models), it is of the utmost importance to decide upon reliability of such potential translation pairs (see research question RQ3). A precision-oriented algorithm for bilingual lexicon extraction tailored to deal with such inherent noise in comparable data should be capable of deciding upon confidence and reliability of potential translation pairs. It should be able to retain only highly confident translation pairs and disregard all the others.

In this chapter, we propose a *generic precision-oriented algorithm* which aims to utilize the output ranked lists coming from an initial model of similarity and their properties in retaining only highly confident translation pairs. The algorithm may be observed as a *post-processing step* which targets to remove

noise propagated from distributional clues in comparable corpora to the output ranked lists of semantically similar words cross-lingually and, consequently, build bilingual lexicons of only highly confident translation pairs.

The algorithm is based on two key assumptions which we "re-apply" to the cross-lingual setting with comparable data: (1) the *symmetry assumption*, and (2) the *one-to-one constraint*. The idea of symmetrization has been borrowed from the symmetrization heuristics introduced for word alignments in statistical machine translation (SMT) [229, 154], where the intersection heuristic is employed for a precision-oriented final alignment. In our setting, it basically means that we take into consideration a translation pair $(w_1^S, w_2^T)$ if and only if, after the symmetrization process and re-ranking of the initial output ranked lists of nearest neighbors, the translation candidate for the source language word $w_1^S$ is the target language word $w_2^T$, and vice versa. The one-to-one constraint targets to match the most confident candidates during the early stages of our precision-oriented algorithm, and then the algorithm excludes the highly confident pair from the search space in further searches. The utility of this constraint for alignment models from parallel corpora has been evaluated by Melamed [197].

In this chapter, we present a case study, and thoroughly describe and evaluate a precision-oriented algorithm which relies on the best scoring TI+Cue model of cross-lingual similarity from sect. 6.3.7 in chapter 6. However, the reader must be aware that the description of the algorithm is rather general and the same reasoning and modeling with only slight modifications may be applied to build the post-processing precision-oriented algorithms, which are built upon all other statistical models of cross-lingual similarity which output the ranked lists of semantically similar words.

The remainder of the chapter is structured as follows. In sect. 7.2 we motivate the main assumptions behind our precision-oriented algorithm. In sect. 7.3 the full algorithm is described. Our experimental setup is listed in sect. 7.4, while sect. 7.5 tests the utility of the proposed algorithm in a series of experiments. Finally, conclusions of the chapter are summarized in sect. 7.6.

## 7.2 Main Modeling Assumptions

This section explains the underlying assumptions of the algorithm: the symmetry assumption and the one-to-one constraint which serve as the backbone of the precision-oriented algorithm.

### 7.2.1 Symmetry Assumption

First, we start with the intuition that the *symmetry assumption* strengthens the confidence of a translation pair. In other words, if the translation candidate for a source language word $w_1^S$ is a target language word $w_2^T$ and, vice versa, the translation candidate of the target word $w_2^T$ is the source word $w_1^S$, and their respective scores in the output ranked lists are above a certain threshold, we can claim that the words $w_1^S$ and $w_2^T$ constitute a *candidate translation pair*. The strict definition of the symmetric relation may also be relaxed. Instead of observing only the translation candidate (i.e., only the first nearest neighbor, see sect. 6.2.1) from the ranked list, we may observe top $M$ candidates from both sides (i.e., source-to-target and target-to-source) and include them in the search space. We may then re-rank the ranked lists comprising the $M$ nearest neighbors, taking into account their associated similarity scores (obtained by our model of similarity) and their respective positions in the ranked lists. The pruning position $M$ may also be referred to as the *search space depth*. The outline of the re-ranking method with the search space comprising the top $M$ nearest neighbors in both directions is provided in alg. 7.1.

We call this symmetrization process the *symmetrizing re-ranking*. It attempts at pushing the correct translation candidate (i.e., the actual cross-lingual synonym) to the top of the new "re-ranked" ranked list, taking into account both the strength of similarities defined through similarity scores in both directions, and positions in ranked lists. Scores $G_{1,i}$ and $G_{2,i}$ aim to balance between positions in the ranked lists and similarity scores, since they reward words which have high similarity scores associated with them, and penalize words if they are found lower in the ranked list. We may also design a thresholded variant of this symmetrizing re-ranking procedure by imposing an extra constraint. After calculating the ranked list $RL(w_1^S)$ for the source language word $w_1^S$, we proceed to alg. 7.1 only if the translation candidate from $RL(w_1^S)$ is assigned a score above a certain threshold $\Delta$. Additionally, in step 3(c), we retain pruned ranked lists $RL_M(w_{1,i}^T)$ of $M$ best scoring source language words only for the target language words $w_{1,i}^T$ for which the first source language translation candidate in their respective ranked list scored above the same threshold $\Delta$. We will call this procedure the *thresholded symmetrizing re-ranking*, and this version will be employed in the final algorithm.

### 7.2.2 One-to-One Constraint

Melamed [197] has already established that many source language words in parallel corpora are non-ambiguous as they tend to translate to only one

**Algorithm 7.1**: SYMMETRIZING RE-RANKING

**Input** : source language word $w_1^S$, ranked list $RL(w_1^S)$ obtained by a model of cross-lingual semantic similarity ;

1: **initialize** an empty "re-ranked" ranked list $RRL(w_1^S) = \{\}$ where a subset of target language words with their recalculated similarity scores will be stored ;

2: **retain** only $M$ best scoring nearest neighbors from $RL(w_1^S)$ along with their respective similarity scores $sim(w_1^S, w_{1,i}^T)$, $sim(w_1^S, w_{1,i}^T) \geq sim(w_1^S, w_{1,i+1}^T)$ $\rightarrow RL_M(w_1^S) = \{w_{1,1}^T, \ldots, w_{1,M}^T\}$ ;

3: **foreach** $w_{1,i}^T \in \{w_{1,1}^T, \ldots, w_{1,M}^T\}$ **do**

- (a): **obtain** $RL(w_{1,i}^T)$ ;
- (b): **retain** only $M$ best scoring nearest neighbors $\rightarrow RL_M(w_{1,i}^T)$ ;
- (c): **if** $w_1^S \in RL_M(w_{1,i}^T)$ **then**
  - **remember**: (1) position $m$, denoting how high in the list $RL_M(w_{1,i}^T)$ the word $w_1^S$ was found, (2) similarity score $sim(w_{1,i}^T, w_1^S)$ ; **calculate**:
  - (i) $G_{1,i} = sim(w_1^S, w_{1,i}^T)/i$ ;
  - (ii) $G_{2,i} = sim(w_{1,i}^T, w_1^S)/m$ ;
  - (iii) $GM_i = \sqrt{G_{1,i} \cdot G_{2,i}}$ ;
  - **add** a tuple $(w_1^S, w_{1,i}^T, GM_i)$ to $RRL(w_1^S)$ ;

4: **if** $RRL(w_1^S)$ *is not empty* **then**

- (a) **sort** the tuples in $RRL(w_1^S)$ in descending order according to their respective $GM_i$ scores ;
- (b) **extract** a translation pair $(w_1^S, w_{1,high}^T)$, where $w_{1,high}^T$ is the best scoring word from $RRL(w_1^S)$ ;

**Output**: sorted list $RRL(w_1^S)$; Translation pair $(w_1^S, w_{1,high}^T)$;

target word. This tendency, although not true in general (see later in chapter 10) is modeled by the *one-to-one constraint*, which constrains each source language word to have at most one translation on the target language side. Melamed's paper reports that this bias leads to a significant positive impact on precision and recall of bilingual lexicon extraction from parallel corpora. This assumption should also be reasonable for many types of comparable corpora such as Wikipedia or news corpora, which are topically aligned or cover similar themes. We will prove that the assumption leads to better precision scores even for bilingual lexicon extraction from such comparable data. The reader has to be aware that the one-to-one constraint is a rather pragmatic heuristic and does not have a strong theoretical background as it completely removes the notion of ambiguity of words from the modeling perspective. However, we believe that introducing this constraint will lead to more help than harm. Without the

one-to-one assumption, similarity scores between source and target language words are calculated independently of each other. We will illustrate the problem arising from the independence assumption with an example.

Suppose that we have an Italian word *arcipelago*, and we would like to detect its correct English translation (*archipelago*). However, after the TI+Cue similarity model is employed, and even after the symmetrizing re-ranking process from the previous step is used, we still acquire an incorrect candidate translation pair *(arcipelago, island)*. Why is that so? The word (*arcipelago*) (and its translation) and the acquired translation (*island*) are semantically very similar, and therefore have similar distributions over latent cross-language topics, but *island* is a much more frequent term. The similarity model such as TI+Cue typically assigns more importance to more frequent candidates, so it will eventually end up learning an *indirect association*.[1] The one-to-one constraint should mitigate the problem of such indirect associations if we design our algorithm in such a way that it learns the most confident *direct associations* first:

1. Learn the correct direct association *(isola, island)*.
2. Remove the words *isola* and *island* from their respective vocabularies.
3. Since *island* is not in the vocabulary, the indirect association between *arcipelago* and *island* is not present any more. The algorithm learns the correct direct association *(arcipelago, archipelago)*.

An illustrative example depicting how both assumptions help boosting accuraccy of learned translation pairs is presented in fig. 7.1. If we retain top $M = 3$ nearest neighbors from both sides and apply both assumptions, the algorithm is able to detect that the correct Dutch-English translation pair is *(abdij, abbey)*. The TI+Cue model without any assumptions would result in an indirect association *(abdij, monastery)*. If only the one-to-one constraint was present, the algorithm would greedily learn the correct direct association *(monastery, klooster)*, remove those words from their respective vocabularies and then again result with another indirect association *(abdij, monk)*. By additionally employing the symmetry assumption with the symmetrizing re-ranking method from sect. 7.2.1, the algorithm correctly learns the translation pair *(abdij, abbey)*. Correct translation pairs *(klooster, monastery)* and *(monnik, monk)* are also obtained. Again here, the pair *(monnik, monk)* would not have been obtained without the introduction of the one-to-one constraint. Knowing all this, it is about time to dive into the final structure of the algorithm.

[1] A direct association, as defined in [197] and already discussed in sect. 5.5.2 in chapter 5, is an association between two words where the two words are indeed mutual translations. Otherwise, it is an indirect association.

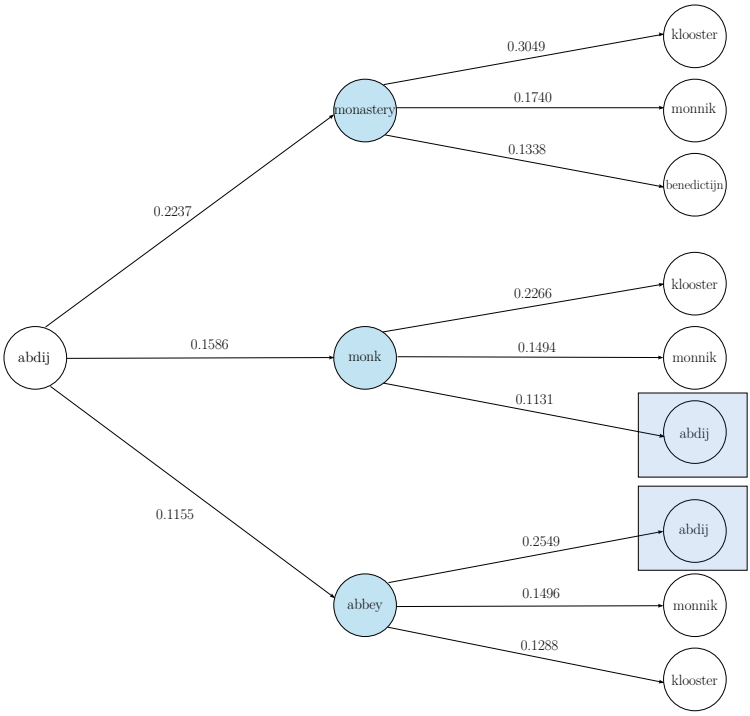



Figure 7.1: An illustrative example depicting the basic advantages of introducing the symmetry assumption and the one-to-one constraint.

## 7.3 Algorithm for Selecting Highly Confident Pairs

### 7.3.1 One-Vocabulary-Pass

First, we present a version of the algorithm with a fixed threshold $\Delta$ which completes only one pass through the complete source language vocabulary $V^S$. Before we start, we need to define several parameters of the algorithm. Let $M_0$ be the initial maximum search space depth for the thresholded symmetrizing re-ranking procedure. Let $M_c$ be the current search space depth, and let $M_{cmax}$ denote the current maximum search depth. For instance, in fig. 7.1, the current depth $M_c$ is 3, while the current maximum depth $M_{cmax}$ might be set to a value higher than 3. The *One-Vocabulary-Pass* algorithm which proceeds with the fixed threshold $\Delta$ is summarized in alg. 7.2. The intuition behind always starting with $M_c = 1$ (step 3(a)) is simple - we are trying to detect a direct association as high as possible in the ranked list. In other words, if the translation candidate for the source word *isola* is the target word *island*, and,

**Algorithm 7.2**: ONE-VOCABULARY-PASS

**initialize** the maximum search space depth: $M_{cmax} := M_0$ ;
**initialize** an empty one-to-one lexicon: $L_o$ ;
**foreach** $w_i^S \in V^S$ **do**
  (a) **set** the current search space depth: $M_c := 1$ ;
  (b) **perform** the thresholded symmetrizing re-ranking procedure (alg. 7.1 and sect. 7.2.1) with the current search space $M_c$ and the threshold $\Delta$ ;
  (c) **if** *a translation pair* $(w_1^S, w_{1,high}^T)$ *is found* **then**
    (1) **remove** words $w_1^S$ and $w_{1,high}^T$ from their respective vocabularies to satisfy the one-to-one constraint:
    $V^S = V^S - \{w_1^S\}$; $V^T = V^T - \{w_{1,high}^T\}$ ;
    (2) **add** the pair $(w_1^S, w_{1,high}^T)$ to the lexicon $L_o$ ;
  **else**
    **if** $M_c < M_{cmax}$ **then**
      $M_c := M_c + 1$ ;
      **go** back to step 3(b) ;
    **else**
      **continue** ;

vice versa, the first translation candidate for the target word *island* is *isola*, we do not need to expand our search depth further, because these two words are the most likely translations. In summary, we will employ the procedure from alg. 7.2 later in an iterative algorithm with a varying threshold and a varying maximum search space depth.

### 7.3.2 The Final Algorithm: SelRel

Let us now define $\Delta_0$ as the initial threshold, which is typically set to a high value in order to determine only the most confident translations in the first passes of the iterative process. Let $\Delta_{min}$ be the threshold at which we stop decreasing the value for this threshold, and start expanding our maximum search space depth for the thresholded symmetrizing re-ranking procedure, and let $dec_\Delta$ be a value for which we decrease the current threshold in each step. Finally, let $M_{max}$ be the final limit for the maximum search space depth, and $M_{cmax}$ is again the current maximum search space depth. The final precision-oriented algorithm which *sel*ects only highly *rel*iable translation pairs called *SelRel* is presented by alg. 7.3.

The parameters of the algorithm model its behavior. Typically, we would like to

**Algorithm 7.3**: SELREL ALGORITHM

1: **initialize** the maximum search space depth $M_{cmax} = M_0$ and the starting threshold $\Delta = \Delta_0$ ;
2: **initialize** an empty lexicon $L_f$ ;
3: **check** the stopping criterion:
**if** $M_{cmax} > M_{max}$ **then**
| **go** to step 9 ;
**else**
| **continue** to step 4 ;
4: **perform** *One-Vocabulary-Pass* from its step 2 (alg. 7.2) with the current values of $\Delta$ and $M_{cmax}$ ;
**obtain** a lexicon $L_o$ ;
5: $L_f := L_f \cup L_o$ ;
6: $L_o := \{\}$ ;
7: **decrease** $\Delta := \Delta - dec_\Delta$ ;
8: **if** $\Delta \geq \Delta_{min}$ **then**
| **go** back to step 4 ;
**else**
| **reset** $\Delta$: $\Delta := \Delta_0$ ;
| **increment** $M_{cmax}$: $M_{cmax} := M_{cmax} + 1$ ;
| **go** back to step 2 ;
9: **return** $L_f$ ;
**Output**: One-to-one bilingual lexicon $L_f$;

set $\Delta_0$ to a high value, and $M_0$ to a low value, which makes our constraints strict and narrows our search space, and consequently, extracts less translation pairs in the first steps of the algorithm, but the set of those translation pairs should be highly confident. Once it is not possible to extract any more pairs with such strict constraints, the algorithm relaxes them by lowering the threshold and expanding the search space by incrementing the maximum search space depth. The algorithm may leave some of the source language words unmatched, which is also dependent on the parameters of the algorithm, but, due to the one-to-one assumption, that scenario will also occur whenever a target vocabulary contains more words than a source vocabulary. The goal of the algorithm is not to find translations for all words, but to select a subset of highly confident translation pairs only.

## 7.4 Experimental Setup

**Training Collections.** Besides the Italian-to-English BLE task used for evaluation in chapter 6, we additionally include the Dutch-to-English BLE task to test whether our precision-oriented algorithm is robust across different language pairs. We use exactly the same dataset for Italian-English as in sect. 6.4.2 with 7,160 Italian nouns and 9,116 English nouns. The Dutch-English training collection comprises 7,612 Dutch-English aligned Wikipedia article pairs augmented with 6,206 Dutch-English document pairs from Europarl [153]. Although Europarl is a parallel corpus, no explicit use is made of sentence-level alignments, and we treat it only as a document-aligned corpus. After the same preprocessing step as in sect. 6.4.2, our final vocabularies consist of 17,754 Dutch nouns and 15,745 English nouns.

**Multilingual Topic Model.** As hinted in sect. 6.5.1, all further experiments with MuPTM-based models of similarity in Part III will rely on the BiLDA model trained with Gibbs sampling and will use the following parameter settings: $K = 2000$, $\alpha = 50/K$, $\beta = 0.01$. A remark for all the following chapters: We are well aware that different hyper-parameter settings [11, 182], might have influence on the quality of learned cross-lingual topics, but that analysis is out of the scope of this thesis.

**Test Data and Ground Truth.** As with Italian-English, we have also created a set of 1,000 ground truth one-to-one translation pairs for Dutch-English following the same procedure (see sect. 6.4.2).

**Evaluation Metrics**. One evaluation metric used is $Acc_1$ already introduced in sect. 6.4.2 (see eq. (6.16)). $Acc_1$ may also be observed as a *Recall* score measuring the recall at pruning position 1. Since the SelRel algorithm does not provide a translation candidate for every single word from the source language vocabulary, we may also define a similar *Precision* score as follows:

$$Precision = \frac{1}{|W_G|}|\{w_i^S : w_i^S \in G \wedge (w_i^S, GTC(w_i^S)) \in L_f\}| \tag{7.1}$$

where $|W_G|$ denotes a set of source language words $w_i^S \in G$ for which the algorithm found a corresponding translation candidate (i.e., the pair $(w_i^S, w_{i,high}^T)$ exists in $L_f$ after alg. 7.3 has been employed). In one experiment (see sect. 7.5.3), we have also computed the $F_\eta$ measure [306], which is computed as follows:

$$F_\eta = (1 + \eta^2)\frac{Precision \cdot Recall}{\eta^2 \cdot Precision + Recall} \tag{7.2}$$

Since our algorithm is precision-oriented, we value precision more than recall. We have set $\eta = 0.5$, which values precision as twice as important as recall.

**Algorithm Parameters.** The parameters of the algorithm have been adjusted on the development set comprising 200 Italian-English translation pairs (see sect. 6.4.2), and have not been further optimized for Dutch-English. The parameters are set to the following values in all experiments, except where noted different: $\Delta_0 = 0.20$, $\Delta_{min} = 0.00$, $dec_\Delta = 0.01$, $M_0 = 3$, and $M_{max} = 10$.

## 7.5 Experiments, Results and Discussion

In this section, we evaluate our precision-oriented algorithm in a series of experiments: (1) We test the utility of our modeling assumptions; (2) We measure how thresholding affects the accuracy of extracted translation pairs; and (3) We build a full precision-oriented lexicon blending translation pairs extracted by SelRel with translation pairs obtained by applying simple language-specific rules.

### 7.5.1 Experiment I: Do Our Modeling Assumptions Help Bilingual Lexicon Extraction?

With this set of ablation experiments, we test whether both the symmetry assumption and the one-to-one constraint are useful in improving quality of the initial one-to-one bilingual lexicon extracted by the TI+Cue similarity model. We compare three different bilingual lexicons: (1) the basic lexicon harvested from TI+Cue ranked lists (*TI+Cue-Basic*) which serves as our baseline, (2) the lexicon obtained after applying the post-processing SelRel algorithm from alg. 7.3, but without the one-to-one constraint (*TI+Cue-Sym*), meaning that if we find a translation pair, we still retain words from the translation pair in their respective vocabularies, (3) the lexicon obtained after applying the one-to-one constraint, but without any symmetrization process (*TI+Cue-One*), and (4) the complete algorithm from alg. 7.3 (*TI+Cue-SelRel*). $Acc_1$ scores for the four lexicon extraction models for both language pairs are provided in tab. 7.1.

Based on these results, it is clearly visible that both modeling assumptions are valid and contribute to better overall scores. Appending the symmetrization process as the post-processing step leads to better bilingual lexicons. Furthermore, we may observe that by additionally imposing the one-to-one constraint, we are able to further increase the $Acc_1$ scores. It is interesting to point out that, although the Dutch-English task uses more comparable data (e.g., it augments the Wikipedia corpus with Europarl), results for that task are slightly lower. The main reason lies in the fact that many Dutch compounds (as Dutch is a

compounding language) cannot be translated by only one word in English. We have also dealt with larger vocabularies for Dutch-English.

| TI+Cue- | **IT→EN** | **NL→EN** |
|---|---|---|
| Basic | 0.597 | 0.446 |
| Sym | 0.612 | 0.469 |
| One | 0.614 | 0.465 |
| SelRel | **0.633** | **0.498** |

Table 7.1: $Acc_1$ scores for 2 language pairs with our 4 BLE algorithms.

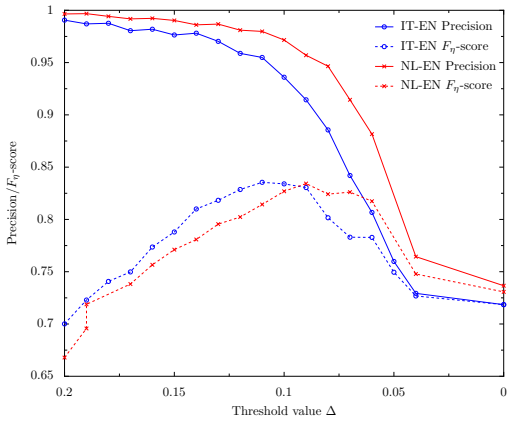



Figure 7.2: Precision and $F_{0.5}$ scores in relation to threshold values.

### 7.5.2 Experiment II: Thresholding and Precision?

The next set of experiments aims at exploring how precision scores (see eq. (7.1)) change while we gradually decrease threshold values $\Delta$. The main goal of these experiments is to detect when to stop with the extraction of translation candidates in order to preserve a bilingual lexicon of only highly confident translation pairs. We fix the maximum search space depth $M_0 = M_{max} = 3$. Fig. 7.2 displays the change of precision in relation to different threshold values, where we start harvesting translations from the threshold $\Delta_0 = 0.2$ down to $\Delta_{min} = 0.0$. Since our goal is to extract as many correct translation pairs as possible, but without decreasing the precision scores, we have also examined what impact this gradual decrease of threshold also has on the number of extracted translations (as reflected in *Recall* and $F_\eta$ scores, see eq. (6.16) and eq. (7.2)). The $F_{0.5}$ scores are also provided in fig. 7.2.

We can observe that the SelRel algorithm retrieves only highly confident translations for both language pairs while the threshold goes down from value 0.2 to 0.1, while precision starts to decrease significantly after the threshold of 0.1. $F_{0.5}$ scores also reach their peaks within that threshold region. We may also observe that larger threshold values (in the [0.15-0.20] region) lead to very confident translation pairs (as reflected in extremely high precision scores above 0.95), but the amount of extracted candidates is extremely low (as reflected in lower $F_{0.5}$ scores). The reader has to be aware that these numbers are specific

to the TI+Cue similarity model. Applying the SelRel post-processing algorithm to some other similarity model would result in a similar behavior, but with different absolute parameter values.

### 7.5.3 Experiment III: Building a Highly Confident Lexicon

Finally, we test how many translation pairs our SelRel algorithm is able to acquire from the entire source vocabulary, with very high reliability of the pairs (i.e., precision of the algorithm) still remaining paramount. If we do not possess any knowledge about a given language pair, we may use only words shared across languages as lexical clues for the construction of a seed lexicon. It often leads to a lower precision lexicon, due to a problem with *false friends*. False friends are pairs of words or phrases in two languages or dialects that look or sound similar, but differ significantly in meaning. Some examples of Italian-English false friends are *pane (bread)-pane*, or *kind (child)-kind* for Dutch-English. For Italian-English, we have found 431 nouns shared between the two languages, of which 350 were correct translations, leading to a precision score of 0.812. As an illustration, if we take the first 431 translation pairs retrieved by the SelRel algorithm, there are 427 correct translation pairs, leading to a precision of 0.9907. Some pairs do not share any orthographic similarities: *(uccello, bird), (tastiera, keyboard), (salute, health), (terremoto, earthquake)*, etc.

Besides the words shared between two languages, following Koehn and Knight [155], we have also employed simple transformation rules for the adoption of words from one language to another. The rules specific to the Italian-English translation process that have been employed are: (Rule-1) if an Italian noun ends in $-ione$, but not in $-zione$, strip the final $e$ to obtain the corresponding English noun. Otherwise, strip the suffix $-zione$, and append $-tion$; (Rule-2) if a noun ends in $-ia$, but not in $-zia$ or $-fia$, replace the suffix $-ia$ with $-y$. If a noun ends in $-zia$, replace the suffix with $-cy$ and if a noun ends in $-fia$, replace it with $-phy$. Similar rules have been introduced for Dutch-English: the suffix $-tie$ is replaced by $-tion$, $-sie$ by $-sion$, and $-teit$ by $-ty$. Finally, we have compared the results of the following automatically constructed lexicons:

(1) A lexicon containing only words shared across languages (LEX-1).

(2) A lexicon containing shared words and translation pairs found by applying the language-specific transformation rules (LEX-2).

(3) A lexicon containing only translation pairs obtained by our SelRel algorithm appended on the TI+Cue similarity model that score above a certain threshold $\Delta$ (that value is $\Delta = 0.10$ according to the findings from sect. 7.5.2) (LEX-SelRel).

| | Italian-English | | | Dutch-English | | |
|---|---|---|---|---|---|---|
| **Lexicon** | Correct | Precision | $F_{0.5}$ | Correct | Precision | $F_{0.5}$ |
| LEX-1 | 350 | 0.812 | 0.188 | 898 | 0.862 | 0.231 |
| LEX-2 | 766 | 0.894 | 0.347 | **1376** | 0.901 | **0.322** |
| LEX-SelRel | **782** | **0.896** | **0.352** | 1106 | **0.956** | 0.278 |
| LEX-1+LEX-SelRel | 1070 | 0.879 | 0.429 | 1860 | 0.908 | 0.396 |
| LEX-R+LEX-SelRel | 1141 | **0.924** | 0.455 | 1507 | **0.964** | 0.350 |
| LEX-2+LEX-SelRel | **1429** | 0.893 | **0.510** | **2261** | 0.922 | **0.451** |

Table 7.2: A comparison of different precision-oriented bilingual lexicons for Italian-English and Dutch-English in terms of the number of correct translation pairs, precision and $F_{0.5}$ scores.

(4) A combination of the lexicons LEX-1 and LEX-SelRel (LEX-1+LEX-SelRel). Non-matching duplicates are resolved by taking the translation pair from LEX-SelRel as the correct one. Note that this lexicon is still completely language pair independent.

(5) A lexicon combining only translation pairs found by applying the language-specific transformation rules and LEX-SelRel (LEX-R+LEX-SelRel).

(6) A combination of the lexicons LEX-2 and LEX-SelRel, where non-matching duplicates are resolved by taking the translation pair from LEX-SelRel (LEX-2+LEX-SelRel).

According to the results from tab. 7.2, we may conclude that adding translation pairs extracted by our SelRel algorithm on top of the TI-Cue similarity model has a major positive impact on both precision and coverage. Obtaining results for two different language pairs proves that the algorithm is generic and applicable to more language pairs. The previous approach relying on work from Koehn and Knight [155] has been outperformed in terms of precision and coverage. Additionally, we have shown that the addition of simple translation rules for languages sharing the same roots might lead to even better scores (LEX-2+LEX-SelRel). However, it is not always possible to rely on such knowledge, and the usefulness of the designed SelRel algorithm should really come to the fore when the algorithm is applied on more distant language pairs which do not share many words and cognates, and word translation rules cannot be easily established. In such cases, without any prior knowledge about the languages involved in a translation process, one is left with the linguistically unbiased LEX-1+LEX-SelRel lexicon, which also displays a promising performance.

## 7.6 Conclusions and Future Work

In this chapter, we have further extended our statistical framework for modeling cross-lingual semantic similarity and bilingual lexicon extraction by presenting a novel precision-oriented algorithm called SelRel, which selects only highly confident translation pairs given the knowledge of ranked lists obtained by an initial similarity model. Put simply, our aim in this chapter was to further work on the solution for research question RQ2, but now also tackling research question RQ3, that is, we wanted to test whether highly confident translation pairs may be extracted from noisy and unstructured comparable data. The precision-oriented algorithm, which can be observed as a post-processing step applied on top of the initial model of similarity, is based on two key assumptions: (1) the symmetry assumption, and (2) the one-to-one constraint. We have empirically proven the utility of these assumptions and have evaluated our algorithm and investigated its properties in a series of experiments. We have shown that the SelRel algorithm is able to produce highly reliable translation pairs, which is especially important when dealing with noisy environments such as comparable corpora without any other lexical clues.

In this chapter, we have presented the effect of the SelRel algorithm applied on top of the TI+Cue similarity model. However, the similar idea, that is, an adjusted version of the same algorithm underpinned by the symmetry assumption and the one-to-one constraint might be applied to other models of similarity, as long as these models provide ranked lists of semantically similar words.

## 7.7 Related Publications

[1] **I. Vulić** and M.-F. Moens. "Detecting highly confident word translations from comparable corpora without any prior knowledge," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France, 23-27 April 2012, pp. 449-459, ACL, 2012.

# 8

# Cross-Lingual Similarity of Words as the Similarity of Their Semantic Word Responses

*Thinking doesn't seem to help very much. The human brain is too high-powered to have many practical uses in this particular universe.*

— Kurt Vonnegut Jr.

## 8.1 Introduction

In this chapter, we further extend our statistical framework for modeling cross-lingual semantic similarity (research question RQ2). We present a new set of shared cross-lingual context features (see sect. 6.2.2) which consequently span a new shared cross-lingual semantic space. The approach to constructing the shared cross-lingual semantic space relies on a paradigm of *semantic word responding* or *free word association*. We borrow that concept from the psychology/cognitive science literature. Semantic word responding addresses a task that requires participants to produce first words that come to their mind that are related to a presented cue word [219, 279].

This new shared cross-lingual semantic space is spanned *by all vocabulary words in the source and the target language*. Each axis in the space denotes a semantic word response. The similarity between two words is then computed as the similarity between the vectors comprising their semantic word responses again

using any of existing *SF*-s. *Two words are considered semantically similar if they are likely to generate similar semantic word responses and assign similar importance to them.*

We utilize a shared semantic space of latent cross-lingual topics learned by a multilingual probabilistic topic model to obtain semantic word responses and quantify the strength of association between any cue word and its responses monolingually and across languages, and, consequently, to build *semantic response feature vectors*. It effectively translates the task of word similarity from the semantic space spanned by latent cross-lingual topics to the semantic space spanned by all vocabulary words in both languages, where the score of each dimension $w_i \in V^S \cup V^T$ for some cue word $w_1^S$ is exactly the strength of the free word association between $w_1^S$ and $w_i$, regardless of the actual language to which $w_i$ belongs.

In order to quantify the strength of free word association or semantic word responding, we apply our Cue model of similarity from sect. 6.3.5 in chapter 6. The prior work in psychology and cognitive science [114, 278] has shown that this model closely mimics the behavior of the human brain when dealing with the task of semantic word responding and produces very good results for that task [114]. In this chapter, we will discuss the model of similarity in more detail and draw direct analogies to the cognitive science literature and the "semantics of the human brain". Following that, we will demonstrate how to utilize the new shared semantic space based on the concept of semantic word responding. In summary, the main contributions of this chapter are as follows:

(i) We propose a new approach to modeling cross-lingual semantic similarity of words based on the similarity of their semantic word responses.

(ii) We present how to estimate and quantify semantic word responses both monolingually and across languages by means of a multilingual probabilistic topic model and analyze the implications and the connections of the estimation with cognitive science.

(iii) We show that the *second-order response-based model of similarity* is more robust and obtains better results for bilingual lexicon extraction (BLE) than the models that operate in the semantic space spanned by latent cross-lingual topics directly. In other words, by performing a transition from the semantic space spanned by latent cross-lingual topics to the semantic space spanned by semantic responses, we build more robust models of semantic similarity.

The following sections first illustrate the intuition behind the response-based approach, demonstrate how to quantify the strength of free word association and then describe our response-based approach to modeling cross-lingual semantic word similarity (sect. 8.2). Following that, we present our evaluation and results

on the BLE task for three language pairs in sect. 8.4. Sect. 8.5 summarizes main conclusions and provides directions for future work.

## 8.2 Modeling Cross-Lingual Word Similarity as the Similarity of Semantic Word Responses

### 8.2.1 The Intuition Behind the Approach

Imagine the following thought experiment. A group of human subjects who have been raised bilingually and are therefore native speakers of two languages $L_S$ and $L_T$, is playing a game of word associations. The game consists of possibly an infinite number of iterations, and each iteration consists of four rounds. In the first round (the *S-S round*), given a word in the language $L_S$, the subject has to generate a list of words in the same language $L_S$ that first occur to her/him as semantic word responses to the given word. The list is in descending order, with more prominent word responses occurring higher in the list. In the second round (the *S-T round*), the subject repeats the procedure, and generates the list of word responses to the same word from $L_S$, but now in the other language $L_T$. The third (the *T-T round*) and the fourth round (the *T-S round*) are similar to the first and the second round, but now a list of word responses in both $L_S$ and $L_T$ has to be generated for some cue word from $L_T$. The process of generating the lists of semantic responses then continues with other cue words and other human subjects.

As the final result, for each word in the source language $L_S$, and each word in the target language $L_T$, we obtain a single list of semantic word responses comprising words in both languages. All lists are sorted in descending order, based on some association score that takes into account both the number of times a word has occurred as an associative response, as well as the position in the list in each round. We can now measure the similarity of any two words, regardless of their corresponding languages, according to the similarity of their corresponding lists that contain their word responses. Words that are equally likely to trigger the same associative responses in the human brain, and moreover assign equal importance to those responses, as provided in the lists of associative responses, are very likely to be closely semantically similar. Additionally, for a given word $w_1^S$ in the source language $L_S$, some word $w_2^T$ in $L_T$ that has the highest similarity score among all words in $L_T$ should be a direct word-to-word translation of $w_1^S$, that is, its translation candidate.

### 8.2.2 Modeling Semantic Responses via Cross-Lingual Topics

Latent cross-lingual topics provide a sound framework to construct a probabilistic model of the aforementioned experiment and to quantify the strength of semantic word associations between words in different languages. To model semantic word responses via the shared space of cross-lingual topics, we have to set a probability mass that quantifies the degree of association. Given two words $w_1, w_2 \in V^S \cup V^T$, a natural way of expressing the *asymmetric semantic association* is by modeling the probability $P(w_2|w_1)$ [114], that is, the probability to generate word $w_2$ as a response given word $w_1$. We may notice that this is exactly our *Cue* model of similarity from sect. 6.3.5 which can be now applied to quantify the strength of semantic word responses both monolingually and cross-lingually, that is, to compute the probability $P(w_2|w_1)$ as follows:

$$Resp(w_1, w_2) = P(w_2|w_1) = \sum_{k=1}^{K} P(w_2|z_k)P(z_k|w_1) \tag{8.1}$$

The probability scores $P(w_2|z_k)$ select words that are highly descriptive for each particular topic. The probability scores $P(z_k|w_1)$ ensure that topics $z_k$ that are semantically relevant to the given word $w_1$ dominate the sum, so the overall high score $Resp(w_1, w_2)$ of the semantic word response is assigned only to highly descriptive words of the semantically related topics. Using the shared space of cross-lingual topics, semantic response scores can be derived for any two words $w_1, w_2 \in V^S \cup V^T$.

The generative model closely resembles the actual process in the human brain - when we generate semantic word responses, we first tend to associate that word with a related semantic/cognitive concept, in this case a latent cross-lingual topic (the factor $P(z_k|w_1)$), and then, after establishing the concept, we output a list of words that we consider the most prominent/descriptive for that concept (words with high scores in the factor $P(w_2|z_k)$), as claimed and proven in [219, 279]. Due to such modeling properties, this model of semantic word responding tends to assign higher association scores to *high frequency words*. It eventually leads to *asymmetric associations/responses*. We have detected that phenomenon both monolingually and across languages. For instance, the first response to Spanish word *mutación (mutation)* is English word *gene*. Other examples include *caldera (boiler)-steam*, *deportista (sportsman)-sport*, *horario (schedule)-hour* or *pescador (fisherman)-fish*. In the opposite association direction, we have detected top responses such as *merchant-comercio (trade)* or *neologism-palabra (word)*. In the monolingual setting, we acquire English pairs such as *songwriter-music*, *discipline-sport*, or Spanish pairs *gripe (flu)-enfermedad (disease)*, *cuenca (basin)-río (river)*, etc.

### 8.2.3 Response-Based Model of Similarity

Eq. (8.1) provides a way to measure the strength of semantic word responses. In order to establish the final similarity between two words, we have to compare their *semantic response vectors*, that is, their semantic response scores over all words in both vocabularies. The final model of word similarity closely mimics our thought experiment:

(1) For each word $w_i^S \in V^S$, we generate probability scores $P(w_j^S|w_i^S)$ for all words $w_j^S \in V^S$ (the S-S rounds). Note that $P(w_i^S|w_i^S)$ is also defined.

(2) For each word $w_i^S \in V^S$, we generate probability scores $P(w_j^T|w_i^S)$, for all words $w_j^T \in V^T$ (the S-T rounds).

(3) For each word $w_i^T \in V^T$, we generate probability scores $P(w_j^T|w_i^T)$ for all words $w_j^T \in V^T$ (the T-T rounds).

(4) For each word $w_i^T \in V^T$, we generate probability scores $P(w_j^S|w_i^T)$ for all words $w_j^S \in V^S$ (the T-S rounds).

Now, each word $w_i \in V^S \cup V^T$ may be represented by a $(|V^S|+|V^T|)$-dimensional context vector $vec(w_i)$ as follows:[1]:

$$vec(w_i) = [P(w_1^S|w_i), \ldots, P(w_{|V^S|}^S|w_i), \ldots, P(w_{|V^T|}^T|w_i)] \tag{8.2}$$

We have created a language-independent cross-lingual semantic space spanned by all vocabulary words in both languages. Each feature corresponds to one word from vocabularies $V^S$ and $V^T$, while the exact score for each feature in the context vector $vec(w_i)$ is precisely the probability that this word/feature will be generated as a semantic word response given the word $w_i$ as a cue word. The degree of similarity between two words is then computed on the basis of similarity between their feature vectors using some of the standard similarity functions (see sect. 6.2.1 in chapter 6 and [167, 43]).

The novel response-based approach to semantic similarity removes the effect of high-frequency words that tend to appear higher in the lists of semantic word responses. Therefore, the real synonyms and translations should occur as top candidates in the lists of similar words obtained by the response-based method. That property may be exploited to identify one-to-one translations across languages and build a bilingual lexicon. Tab. 8.1 presents an illustration of this intuition.

[1] We assume that the two sets $V^S$ and $V^T$ are disjunct. It means that, for instance, Spanish word *pie (foot)* from $V^S$ and English word *pie* from $V^T$ are treated as two different word types. In that case, it holds $|V^S \cup V^T| = |V^S| + |V^T|$.

| Semantic responses | | | | | | Final Sim. |
|---|---|---|---|---|---|---|
| dramaturgo (ES) | | play (EN) | | playwright (EN) | | dramaturgo (ES) |
| obra (play) | .101 | play | .142 | play | .122 | playwright |
| escritor (writer) | .083 | obra (play) | .111 | escritor (writer) | .087 | dramatist |
| play | .066 | player | .033 | obra (play) | .073 | tragedy |
| writer | .050 | escena (scene) | .031 | writer | .060 | play |
| poet | .047 | jugador (player) | .026 | poeta (poet) | .055 | essayist |
| autor (author) | .041 | adaptation | .025 | poet | .053 | novelist |
| poeta (poet) | .039 | stage | .024 | autor (author) | .046 | drama |
| teatro (theatre) | .030 | game | .022 | teatro (theatre) | .043 | tragedian |
| drama | .026 | juego (game) | .021 | tragedy | .031 | satirist |
| contribution | .025 | teatro (theatre) | .019 | drama | .026 | writer |

Table 8.1: An example of top 10 semantic word responses and the final response-based similarity (last column) for a selection of Spanish and English words. The responses are estimated from Spanish-English Wikipedia data by BiLDA.

**Example.** We can observe several interesting phenomena in tab. 8.1: (1) High-frequency words tend to appear higher in the lists of semantic responses (e.g., *play* and *obra* for all three words); (2) Due to the modeling properties that give preference to high-frequency words (sect. 8.2.2), a word might not generate itself as the top semantic response (e.g., *playwright-play*); (3) Both source and target language words occur as the top responses in the lists; (4) Although *play* is the top semantic response in English for both *dramaturgo* and *playwright*, its list of top semantic responses is less similar to the lists of those two words; (5) Although the English word *playwright* does not appear in the top 10 semantic responses to *dramaturgo*, and *dramaturgo* does not appear in the top 10 responses to *playwright*, the more robust response-based similarity method detects that the two words are actually very similar based on their lists of responses; (6) *dramaturgo* and *playwright* have very similar lists of semantic responses which ultimately leads to detecting that *playwright* is the most semantically similar word to *dramaturgo* across the two languages (the last column), that is, they are direct one-to-one translations of each other; (7) Another English word *dramatist* very similar to Spanish *dramaturgo* is also pushed higher in the final list, although it is not found in the list of top semantic responses to *dramaturgo*.

## 8.3 Experimental Setup

**Training Collections.** We evaluate our models of cross-lingual semantic similarity in the BLE task for three language pairs: (i) Italian-English (IT-EN), Dutch-English (NL-EN), and Spanish-English (ES-EN). We again train BiLDA on aligned Wikipedia article pairs for all three language pairs (see sect. 3.2.2 in chapter 3). Additionally, for Dutch-English as in chapter 7 (sect. 7.4), we

augment the Wikipedia data with a set of aligned Europarl documents. After the same preprocessing steps as in previous chapters (sect. 6.4.2 and sect. 7.4), the final vocabularies are as follows: (1) 7,160 Italian nouns and 9,116 English nouns for Italian-English, (2) 9,439 Spanish nouns and 12,945 English nouns for Spanish-English, (3) 9,172 Dutch nouns, 12,247 English nouns for Dutch-English trained on Wikipedia (Wiki), and (4) 17,754 Dutch nouns and 15,745 English nouns for Dutch-English trained on Wikipedia plus Europarl (Wiki+EP).

**Multilingual Topic Model**. As before in chapter 7 (see sect. 7.4), we train the BiLDA model with Gibbs sampling where $K = 2000$, $\alpha = 50/K$, $\beta = 0.01$.

**Evaluation Metrics**. We again report $Acc_M$ ($Acc_1$ and $Acc_{10}$) scores and $MRR$ scores (see sect. 6.4.2) in the BLE task.

**Test Data and Ground Truth.** As with Italian-English and Dutch-English, we have also created a set of 1,000 ground truth one-to-one translation pairs for Spanish-English following the same procedure (see sect. 6.4.2 and sect. 7.4).

**Compared Models.** We evaluate and compare the following models of cross-lingual similarity in all our experimental runs:

(1) The Cue similarity model from sect. 6.3.5, which regards the lists of semantic word responses across languages obtained by eq. (8.1) directly as the ranked lists of semantically similar words (Cue-Direct).

(2) The TI+Cue similarity model from sect. 6.3.7 which was the best scoring model in chapter 6 without topic space pruning. This model operates in the shared semantic space of latent cross-lingual concepts/topics directly (TI+Cue).

(3) Our BC model from sect. 6.3.4 which was the best scoring model in chapter 6 after the topic space pruning procedure has been employed. This model works with word representations by means of $K$-dimensional context vectors comprising conditional topic probability scores (BC-Topics).

(4) The response-based similarity model relying on the representations of words described in sect. 8.2.3. We again use the Bhattacharyya coefficient (BC) as the similarity function, but now on these $(|V^S| + |V^T|)$-dimensional context vectors in the semantic space spanned by all words in both vocabularies that represent semantic word responses (BC-Responses).

As hinted in sect. 6.5.2, we perform topic space pruning with $K' = 100$ for BC-Topics. We may employ exactly the same pruning procedure on the response-based word representations (see eq. (8.2)). Therefore, when computing the final similarity scores with BC-Responses, we retain only top 2,000 semantic responses for each test word, and compute the scores in the pruned space.

| **Pair:** | **IT-EN** | | | **ES-EN** | | |
|---|---|---|---|---|---|---|
| Model | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ |
| Cue-Direct | 0.501 | 0.576 | 0.740 | 0.332 | 0.437 | 0.675 |
| TI+Cue | 0.597 | 0.702 | **0.897** | 0.429 | 0.569 | 0.828 |
| BC-Topics | 0.578 | 0.667 | 0.834 | 0.433 | 0.576 | 0.843 |
| BC-Responses | **0.622** | **0.729** | 0.882 | **0.517** | **0.635** | **0.891** |

Table 8.2: Results on the BLE task. Language pairs are Italian-English and Spanish-English.

| **Pair:** | **NL-EN** | | (Wiki) | **NL-EN** | | (Wiki+EP) |
|---|---|---|---|---|---|---|
| Model | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ |
| Cue-Direct | 0.186 | 0.254 | 0.423 | 0.344 | 0.450 | 0.652 |
| TI+Cue | 0.225 | 0.296 | 0.459 | 0.446 | 0.569 | 0.808 |
| BC-Topics | **0.237** | 0.314 | 0.489 | 0.534 | 0.630 | 0.836 |
| BC-Responses | 0.236 | **0.320** | **0.511** | **0.574** | **0.653** | **0.864** |

Table 8.3: Results on the BLE task for Dutch-English, with different corpora used for the estimation of semantic word responses.

## 8.4 Experiments, Results and Discussion

Tab. 8.2 displays the performance of each compared model on the BLE task for Spanish-English and Italian-English, while tab. 8.3 shows their performance for Dutch-English emphasizing the difference in scores when we induce semantic responses from different corpora, that is, when the multilingual topic model is trained on Wiki and Wiki+EP. In addition, example lists of semantically similar words over all three language pairs are shown in tab. 8.4. Based on these results, we are able to derive several conclusions:

(i) BC-Responses performs consistently better than the other three models over all corpora and all language pairs. It is more robust and is able to find some translation pairs omitted by the other methods (see also tab. 8.5). The overall quality of the cross-lingual word similarities and lexicons extracted by the response-based model is dependent on the quality of the estimated semantic response vectors. The quality of these vectors is of course further dependent on the quality of the multilingual training data. For instance, for Dutch-English, we may observe a rather spectacular increase in overall scores (the tests are performed over the same set of 1,000 words) when we augment Wikipedia data with Europarl (compare the scores for Wiki and Wiki+EP in tab. 8.3).

| **Italian-English (IT-EN)** | | | **Spanish-English (ES-EN)** | | | **Dutch-English (NL-EN)** | | |
|---|---|---|---|---|---|---|---|---|
| (1) **affresco** (fresco) | (2) **spigolo** (edge) | (3) **coppa** (cup) | (1) **caza** (hunting) | (2) **discurso** (speech) | (3) **comprador** (buyer) | (1) **behoud** (conservation) | (2) **schroef** (screw) | (3) **spar** (fir) |
| *fresco* | polyhedron | club | *hunting* | rhetoric | purchase | *conservation* | socket | conifer |
| mural | polygon | competition | hunt | oration | seller | preservation | wire | pine |
| nave | vertices | final | hunter | *speech* | tariff | heritage | wrap | firewood |
| wall | diagonal | champion | hound | discourse | market | diversity | wrench | seedling |
| testimonial | *edge* | football | safari | dialectic | bidding | emphasis | *screw* | weevil |
| apse | vertex | trophy | huntsman | rhetorician | auction | consequence | pin | chestnut |
| rediscovery | binomial | team | wildlife | oratory | bid | danger | fastener | acorn |
| draughtsman | solid | relegation | animal | wisdom | microeconomics | contribution | torque | girth |
| ceiling | graph | tournament | ungulate | oration | trade | decline | pipe | lumber |
| palace | modifier | soccer | chase | persuasion | listing | framework | routing | bark |

Table 8.4: Example lists of top 10 semantically similar words across all three language pairs according to our BC-Responses similarity model, where the correct translation word is: (column 1) found as the most similar word, (2) contained lower in the list, and (3) not found in the top 10 words in the ranked list.

(ii) A transition from a semantic space spanned by latent cross-lingual topics to a semantic space spanned by vocabulary words leads to better results over all corpora and language pairs. The effect of the transition is best reflected in the scores for BC-Topics and BC-Responses, which are conceptually the same models of similarity, as they harness the same similarity function, but operate in two different semantic spaces. The difference is less visible when using training data of lesser quality (the scores for NL-EN on Wiki). Moreover, since the shared space of cross-lingual topics is used to obtain and quantify semantic word responses, the quality of learned cross-lingual topics influences the quality of semantic word responses. If the semantic coherence of the cross-lingual topical space (see sect. 4.5 in chapter 4) is unsatisfying, the response-based model is unable to generate good semantic response vectors, and ultimately unable to correctly identify semantically similar words across languages.

| **IT-EN** | **ES-EN** | **NL-EN** |
|---|---|---|
| direttore-director | flauta-flute | kustlijn-coastline |
| radice-root | eficacia-efficacy | begrafenis-funeral |
| sintomo-symptom | empleo-employment | mengsel-mixture |
| perdita-loss | descubierta-discovery | lijm-glue |
| danno-damage | desalojo-eviction | kijker-viewer |
| battaglione-battalion | miedo-fear | oppervlak-surface |
| nozione-notion | distribuidor-distributor | lek-leak |

Table 8.5: Example translation pairs found by BC-Responses, but missed by the other three compared models of similarity.

(iii) Due to its modeling properties that assign more importance to high-frequency words, Cue-Direct produces reasonable results in the BLE task only for high-frequency words. Although eq. (8.1) models the concept of semantic word responding in a sound way [114], using the semantic word responses directly is not suitable for the actual BLE task.

(iv) Unlike [155, 117], our response-based model of similarity again does not rely on any orthographic features such as cognates or words shared across languages. It is a pure statistical method that only relies on word distributions over a multilingual corpus. Based on these distributions, it performs the initial shallow semantic analysis of the corpus by means of a multilingual probabilistic model. The method then builds, via the concept of semantic word responding, a language-independent semantic space spanned by all vocabulary words/responses in both languages. That makes the method portable to distant language pairs. However, for similar languages, including more evidence such as orthographic clues might lead to further increase in scores, but we leave that for future work.

## 8.5 Conclusions and Future Work

In this chapter, we have presented and described another extension of our statistical framework for modeling cross-lingual word similarity. This chapter continues the research thread ignited by research question RQ2 - "Is it possible to automatically build statistical data-driven models of cross-lingual semantic similarity (i.e., addressing the *problem of meaning*) a variety of language pairs, without any prior knowledge about language pairs, and without any high-quality external resource (e.g., a hand-crafted domain-specific bilingual lexicon)?". Moreover, we still operate in the noisy setting dealing with comparable data (see again research question RQ5). We have proposed a second-order model of similarity that relies on the paradigm of semantic word responding previously defined in cognitive science. Compared to the original models of similarity proposed in chapter 6, the proposed approach is more robust (but also more complex and computationally expensive). It again does not make any additional language-pair dependent assumptions (e.g., it does not rely on a seed lexicon, orthographic clues or predefined concept categories). That effectively makes it applicable to more language pairs besides the three language pairs used for testing in this chapter. We have presented a new shared cross-lingual semantic space spanned by all vocabulary words in both languages and have demonstrated the utility of this new semantic space. Our experiments on the task of bilingual lexicon extraction for a variety of language pairs have proven that the response-based model of similarity is more robust and outperforms the methods that operate in the semantic space of latent cross-lingual topics.

One line of future work has already been mentioned - we may port our response-model to more language pairs and combine it with other features (e.g., orthographic clues). Another line of future work will be discussed in the following chapter - we may use the initial output of this response-model to move towards even better results in the BLE task by bootstraping another type of a shared cross-lingual semantic space (more to come in the upcoming chapter). Other options for future work include building other models of similarity relying on other similarity functions such as those discussed in chapter 6, or building context-sensitive models of similarity at the word token level using the semantic space of semantic word responses (see chapter 10).

At this point, we would also stress another, more general line of future work which potentials fall beyond the scope of this chapter and this thesis. Namely, recent studies (e.g. [273, 120, 88, 285]) have shown that multilingual representations of words and phrases in shared cross-lingual semantic spaces may prove as an extra source of knowledge even when dealing with monolingual tasks such as word sense disambiguation (e.g., [88, 285]) or measuring monolingual semantic similarity (e.g., [120]). For instance, a straightforward utilization of multilingual

representations monolingually is to tackle the problems of polysemy (i.e., we may detect that two occurrences of the same word type in language $L_S$ represent two different meanings because they exhibit two completely unrelated translations in language $L_T$) and synonymy (i.e., we may detect that two different word types in $L_S$ are semantically similar because they exhibit the same translation in $L_T$). One line of future work may further investigate how to exploit these multilingual representations and shared cross-lingual semantic spaces to improve over sole monolingual representations in monolingual tasks.

## 8.6 Related Publications

[1] **I. Vulić** and M.-F. Moens. "Cross-lingual semantic similarity of words as the similarity of their semantic word responses," in *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, Georgia, USA, 9-15 June 2013, pp. 106-116, ACL, 2013.

# 9

# Bootstrapping Cross-Lingual Vector Spaces (from Almost Nothing)

*Art is making something out of nothing and selling it.*
— Frank Zappa

## 9.1 Introduction

In previous chapters, we have introduced a fully corpus-based framework for computing cross-lingual semantic similarity which relies on a cross-lingual vector space spanned by latent cross-lingual concepts directly (chapter 6) or a cross-lingual vector space spanned by MuPTM-induced semantic word responses (chapter 8). However, the standard way of building a cross-lingual vector space in the relevant literature is to utilize *bilingual lexicon entries* from an existing lexicon [92, 102] as dimensions of the space. But these methods presuppose that there exist readily available bilingual lexicons (which are either hand-crafted or extracted from parallel data) which are then used to induce bilingual lexicons! In order to circumvent this issue, one line of recent work aims to *bootstrap high-quality cross-lingual vector spaces from a small initial seed lexicon*. The seed lexicon is constructed by harvesting identically or similarly spelled words across languages [155, 234], and it spans the initial cross-lingual vector space. *The space is then gradually enriched with new dimensions/axes during the bootstrapping procedure*. The bootstrapping process has already proven its validity in inducing bilingual lexicons for closely similar languages such as Spanish-Portuguese or

Croatian-Slovene [89], but it still lacks further generalization to more distant language pairs.

In this chapter, we tackle the issue of getting caught in that vicious cycle where one needs a lexicon to induce a lexicon, and provide a further generalization of the bootstrapping approaches to inducing cross-lingual semantic similarity. We investigate the utility of bilingual lexicon entries as shared dimensions of the cross-lingual vector space (as opposed to latent cross-lingual topics, see sect. 6.2.2), but propose a new bootstrapping approach to constructing such shared cross-lingual vector spaces which is completely corpus-based and does not require any additional translation resource. We show that selected highly reliable translation pairs from the output pairs obtained by our initial corpus-based models of similarity (chapters 6, 7 and 8) may be employed to commence a bootstrapping procedure. The main goal of the chapter is to shed new light on the bootstrapping approaches to modeling cross-lingual semantic similarity and bilingual lexicon extraction. We show how to construct a language pair agnostic bootstrapping method that is able to build high-quality cross-lingual vector spaces that consequently lead to high-quality bilingual lexicons for more distant language pairs where orthographic similarity is not sufficient to seed cross-lingual vector spaces. We aim to answer the following key questions:

(i) How to seed cross-lingual vector spaces besides using only orthographically similar words?

(ii) Is it better to seed cross-lingual vector spaces with translation pairs/dimensions that are frequent in the corpus, and does the frequency matter at all? Does the size of the initial seed lexicon matter?

(iii) How to enrich cross-lingual vector spaces with only highly reliable dimensions in order to prevent semantic drift?

With respect to these questions, the main contributions of this chapter are:

(i) We present a complete overview of the framework of bootstrapping cross-lingual vector spaces from non-parallel data without any additional resources. We dissect the bootstrapping process and describe all its key components: (1) starting point or seed lexicon, (2) confidence estimation and selection of new dimensions of the space, and (3) convergence.

(ii) We introduce a new way of seeding the bootstrapping procedure that does not rely on any orthographic clues and that yields cross-lingual vector spaces of higher quality. We analyze the impact of different seed lexicons on the quality of induced cross-lingual vector spaces. We also introduce various confidence estimation functions for bootstrapping and analyze their influence on the quality of cross-lingual vector spaces.

(iii) We show that in the setting without any external translation resources, our bootstrapping approach yields lexicons that outperform the previously best performing corpus-based BLE methods (from chapter 6 and chapter 8) on our test datasets for two language pairs.

The remainder of the chapter is structured as follows. First, in sect. 9.2 we present a complete overview of the bootstrapping approach to inducing cross-lingual vector spaces, covering initialization and updating of the space. Following that, we briefly discuss our experimental setup in sect. 9.3. while sect. 9.4 provides results and discussion. Conclusions and future work paths are summarized in sect. 9.5.

## 9.2 Bootstrapping Cross-lingual Vector Spaces: A Complete Overview

This section presents the complete bootstrapping procedure that starts with an initial seed lexicon which spans the initial cross-lingual vector space, while as the output in each iteration of the procedure it produces an updated cross-lingual vector space that can be used to extract a bilingual lexicon.

### 9.2.1 General Framework for Bootstrapping

We again assume that we are solely in possession of a (non-parallel) bilingual corpus $\mathcal{C}$ that is composed of a sub-corpus $\mathcal{C}_S$ given in the source language $L_S$, and a sub-corpus $\mathcal{C}_T$ in the target language $L_T$. The goal is to build a cross-lingual vector space using only corpus $\mathcal{C}$.

**Assumption 1.** *Dimensions of the cross-lingual vector space are one-to-one word translation pairs.* For instance, dimensions of a Spanish-English space are pairs like *(perro, dog)*, *(ciencia, science)*, etc. The *one-to-one constraint* [197] (see also chapter 7), although not valid in general, simplifies the construction of the bootstrapping procedure. $\mathcal{T}$ denotes the set of translation pairs that are the dimensions of the space.

**Computing cross-lingual word similarity in a cross-lingual vector space.** Now, assume that our cross-lingual vector space consists of $N$ one-to-one word translation pairs from $\mathcal{T}$, $c_k = (c_k^S, c_k^T), k = 1, \ldots, N$. For each word $w_i^S \in V^S$, we compute the similarity of that word with each word $w_j^T \in V^T$ by computing the similarity between their context vectors $vec(w_i^S)$ and $vec(w_j^T)$, which are actually their representations in the $N$-dimensional cross-lingual

vector space (see sect. 6.2.1). The cross-lingual similarity is then computed following the standard procedure [93, 102]:

1. For each source word $w_i^S \in V^S$, build its $N$-dimensional context vector $vec(w_i^S)$ that consists of association scores $sc_i^S(c_k^S)$, that is, we compute the strength of association with the "source" part of each dimension $c_k$ that constitutes the $N$-dimensional cross-lingual space. The association is dependent on the co-occurrence of $w_i^S$ and $c_k^S$ in a predefined context. Various functions such as the log-likelihood ratio (LLR) [247, 134], TF-IDF [93], or pointwise mutual information (PMI) [40, 269] are typically used as *weighting functions* to quantify the strength of the association (see sect. 6.2.1).

2. Repeat step (1) for each target word $w_j^T \in V^T$ and build context vectors $vec(w_j^T)$ which consist of scores $sc_j^T(c_k^T)$.

3. Since $c_k^S$ and $c_k^T$ address the same dimension $c_k$ in the cross-lingual vector space for each $k = 1, \ldots, N$, we are able to compute the similarity between $vec(w_i^S)$ and $vec(w_j^T)$ using any similarity measure such as the Jaccard index, the Kullback-Leibler or the Jensen-Shannon divergence, the cosine similarity, or others (see again sect. 6.2.1).

**Bootstrapping.** The key idea of the bootstrapping approach relies on an insight that *highly reliable translation pairs* $(w_1^S, w_2^T)$ which are encountered using the $N$-dimensional cross-lingual vector space might be added as new dimensions of the space. By adding these new dimensions, it might be possible

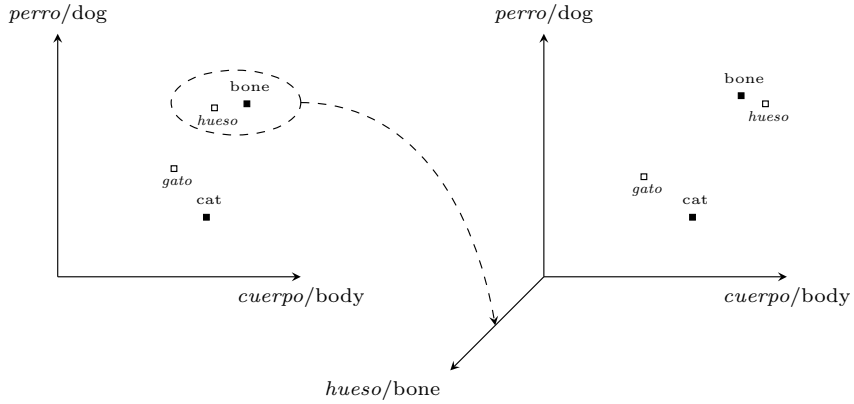



Figure 9.1: An illustration of the bootstrapping approach to constructing a shared Spanish-English cross-lingual vector space.

to extract more highly reliable translation pairs that were previously not used as dimensions of the space, and the iterative procedure repeats until no new dimensions are found. An illustration of this idea is presented in fig. 9.1, where it is displayed how a new dimension $(hueso, bone)$ is added to a Spanish-English cross-lingual vector space which then increases its number of dimensions from two to three. Learning more dimensions of the cross-lingual vector space should intuitively boost expressiveness of the space.

The induced cross-lingual vector space may then be observed as a bilingual lexicon *per se*, but it may also be used to find translation equivalents for other words which are not used to span the space.

**Algorithm 9.1**: BOOTSTRAPPING A CROSS-LINGUAL VECTOR SPACE

**Input** : bilingual corpus $\mathcal{C} = \mathcal{C}_S \cup \mathcal{C}_T$
**initialize**: (1) **obtain** a one-to-one seed lexicon; the entries from the lexicon are initial dimensions of the space: $\mathcal{T}_0$; (2) $s = 0$ ;
**bootstrap**:
**repeat**
1: **foreach** $w_i^S \in V^S$ **do**: **compute** $RL(w_i^S)$ using $\mathcal{T}_s$ ;
2: **foreach** $w_j^T \in V^T$ **do**: **compute** $RL(w_j^T)$ using $\mathcal{T}_s$ ;
3: **foreach** $w_i^S \in V^S$ and $w_j^T \in V^T$ **do**: **score** each translation pair $(w_i^S, TC(w_i^S))$ and $(TC(w_j^T), w_j^T)$ and **add** them to a *pool of candidate dimensions* ;
4: **choose** *the best candidates* from the pool and **add** them as new dimensions: $\mathcal{T}_{s+1} \leftarrow \mathcal{T}_s \cup \{best\}$ ;
5: **resolve** collisions in $\mathcal{T}_{s+1}$ ;
6: $s \leftarrow s + 1$ ;
**until** *no new dimensions are found (convergence)* ;
**Output**: one-to-one translation pairs $\rightarrow$ dimensions of a cross-lingual vector space: $\mathcal{T}_{final}$

The overview of the procedure as given by alg. 9.1 reveals these crucial points in the procedure: (Q1) How to provide initial dimensions of the space? (the initialization step); (Q2) How to score each translation pair, estimate their confidence, and how to choose the best candidates from the pool of candidates (step 3 and step 4), and (Q3) How to resolve potential collisions that violate the one-to-one constraint? (step 5). We will discuss Q1 and Q2 in more detail later, while we resolve Q3 following a simple heuristic as follows:

**Assumption 2.** *In case of collision, dimensions/pairs that are found at later stages of the bootstrapping process overwrite previous dimensions.*

The intuition here is that we expect for the quality of the space to increase at each stage of the bootstrapping process, and newer translation pairs should be more confident than the older ones. For instance, if 2 out of $N$ dimensions of a Spanish-English cross-lingual space are pairs *(piedra,wall)* and *(tapia,stone)*, but then if during the bootstrapping process we extract a new candidate pair *(piedra,stone)*, we will delete the former two dimensions and add the latter.

### 9.2.2 Initializing Cross-Lingual Vector Spaces

Seeding or initializing a bootstrapping procedure is often a critical step regardless of the actual task we aim to solve with the bootstrapping procedure [194, 157], and it decides whether the complete process will end as a success or a failure. However, Peirsman and Padó [235] argue that the initialization step is not crucial when dealing with bootstrapping cross-lingual vector spaces. Here, we present two different strategies of initializing the cross-lingual vector space.

**Identical Words and Cognates.** Previous work relies exclusively on identical and similarly spelled words to build the initial set of dimensions $\mathcal{T}_0$ [155, 234, 89]. This strategy yields promising results for closely similar language pairs, but is of limited use for other language pairs.

**High-Frequency Seeds.** Another problem with using only identical words and cognates as seeds lies in the fact that many of them might be infrequent in the corpus, and as a consequence the expressiveness of a cross-lingual vector space might be limited. On the other hand, high-frequency words offer a lot of evidence in the corpus that could be exploited in the bootstrapping approach. In order to induce initial translation pairs, we rely on our framework for modeling cross-lingual semantic similarity based on latent cross-lingual topics from chapters 6 and 8. We can simply construct the initial seed lexicon from the output of our MuPTM-based models of similarity as follows:

(1) Train a multilingual topic model on the corpus.

(2) Obtain one-to-one translation pairs using any of the MuPTM-based models of cross-lingual similarity (see [312, 314], and chapters 6 and 8).

(3) Retain only *symmetric* translation pairs. This step ensures that only highly confident pairs are used as seed translation pairs.

(4) Rank translation pairs according to their frequency in the corpus and use a subset of the most frequent symmetric pairs as seeds.

### 9.2.3 Estimating Confidence of New Dimensions

Another crucial step in the bootstrapping procedure is the estimation of confidence in a translation pair/candidate dimension. Errors in the early stages of the procedure may negatively affect the learning process and even cause *semantic drift* [254, 54, 194]. A semantic drift in this context denotes errors with negative feedback during the bootstrapping approach, that is, by introducing erroneous dimensions of the space during the bootstrapping process, we might eventually end up learning more and more erroneous dimensions, and the complete process might result in a failure. We therefore impose the constraint which requires translation pairs to be *symmetric* in order to qualify as potential new dimensions of the space. In other words, given the current set of dimensions $\mathcal{T}_s$, a translation pair $(w_1^S, w_2^T)$ has a possibility to be chosen as a new dimension from the pool of candidate dimensions if and only if it holds: $TC(w_1^S) = w_2^T$ and $TC(w_2^T) = w_1^S$. This *symmetry constraint* (see also sect. 7.2.1 in chapter 7) should ensure a relative reliability of translation pairs.

In each iteration of the bootstrapping process, we may add all symmetric pairs from the pool of candidates as new dimensions of the space, or we could impose additional selection criteria that quantify the degree of confidence in translation pairs. We are then able to rank the symmetric candidate translation pairs in the pool of candidates according to their confidence scores (step 3 of alg. 9.1), and choose only the best $B$ candidates from the pool in each iteration (step 4) as done in [288, 194, 132]. By picking only a subset of the $B$ most confident candidates in each iteration, we hope to further prevent a possibility of semantic drift, i.e., "poisoning" the bootstrapping process that might happen if we include incorrect translation pairs as dimensions of the space.

In this chapter, we propose and investigate three different confidence estimation functions:

(1) **Absolute similarity score.** Confidence $CF(w_1^S, TC(w_1^S))$ of a translation pair is simply the absolute similarity value $sim(w_1^S, TC(w_1^S))$.

(2) **M-Best confidence function.** It contrasts the score of the translation candidate with the average score over the first $M$ most similar words in the ranked list. The larger the difference, the more confidence we have in the translation candidate. Given a word $w_1^S \in V^S$ and a pruned ranked list $RL_M(w_1^S)$, the average score of the best $M$ words is computed as:

$$sim_M(w_1^S) = \frac{1}{M} \sum_{w_j^T \in RL_M(w_1^S)} sim(w_1^S, w_j^T) \tag{9.1}$$

The final confidence score is then:

$$CF(w_1^S, TC(w_1^S)) = sim(w_1^S, TC(w_1^S)) - sim_M(w_1^S) \tag{9.2}$$

(3) **Entropy-based confidence function.** We adapt the well-known entropy-based confidence [271, 294] to this particular task. First, we need to define a distribution over all $w_j^T \in V^T$ with scores:

$$P(w_j^T|w_1^S) = \frac{e^{sim(w_1^S, w_j^T)}}{\sum_{w_l^T \in V^T} e^{sim(w_1^S, w_l^T)}} \tag{9.3}$$

The confidence function is then calculated as minus the entropy of the distribution:

$$CF(w_1^S, TC(w_1^S)) = \sum_{w_l^T \in V^T} P(w_l^T|w_1^S) \log P(w_l^T|w_1^S) \tag{9.4}$$

A symmetrized version of the confidence functions is computed as the geometric mean of source-to-target and target-to-source confidence scores.

## 9.3 Experimental Setup

**Training Collections, Test Data, Ground Truth, Evaluation Metrics.** We investigate our bootstrapping approach on the bilingual lexicon extraction (BLE) task for two language pairs: Spanish-English (ES-EN) and Italian-English (IT-EN). We work with the same bilingual Wikipedia data and employ the same preprocessing steps as before (see sect. 8.3). Our ground truth and evaluation metrics ($Acc_M$ and $MRR$) are also left unchanged. Moreover, the BiLDA model is trained with the same parameters as in sect. 8.3.

In chapter 8 we have also worked with Dutch-English (NL-EN), but we have decided to leave out the results obtained for that language pair for the sake of clarity of the presentation, due to space constraints, a high similarity between the two languages, and the fact that the results obtained for that language pair are qualitatively and quantitatively similar to the results we report for ES-EN and IT-EN. Hence including the results for NL-EN would not contribute to this chapter with any new important insight and conclusion.

**Building Initial Seed Lexicons.** To produce the lists of one-to-one translation pairs that are used as seeds for the bootstrapping approach (see sect. 9.2.2), we experiment with the BC-Topics and the BC-Responses models

of cross-lingual similarity from chapter 8 (see sect. 8.3), which are the MuPTM-based models of cross-lingual semantic similarity that obtain the best results in the BLE task on these datasets. The parameters of these models are exactly the same as in chapter 8. These two models also serve as our *baseline models*, and our goal is to test whether we are able to obtain bilingual lexicons of higher quality using bootstrapping that starts from the output of these models.

**Weighting and Similarity Functions.** We have experimented with different families of weighting functions (e.g., PMI, LLR, TF-IDF, chi-square) and similarity functions (e.g., cosine, Dice, Kullback-Leibler, Jensen-Shannon) [167, 299]. In this chapter, we present results obtained by *positive pointwise mutual information* (PPMI) [225] as a weighting function, which is a standard choice in vector space semantics [299], and (combined with cosine) yields the best results over a group of semantic tasks according to [40]. The PPMI score for some source word $w_i^S \in V^S$ and some context feature $c_k$ given an $N$-dimensional cross-lingual vector space with features/dimensions $c_k = (c_k^S, c_k^T), k = 1, \ldots, N$, is computed as follows [299]:

$$p_{ik}^S = \frac{C(w_i^S, c_k^S)}{\sum_{w_j^S \in V^S} \sum_{n=1}^{N} C(w_j^S, c_n^S)} \tag{9.5}$$

$$p_{i*}^S = \frac{\sum_{n=1}^{N} C(w_i^S, c_n^S)}{\sum_{w_j^S \in V^S} \sum_{n=1}^{N} C(w_j^S, c_n^S)} \quad p_{*k}^S = \frac{\sum_{w_j^S \in V^S} C(w_j^S, c_k^S)}{\sum_{w_j^S \in V^S} \sum_{n=1}^{N} C(w_j^S, c_n^S)} \tag{9.6}$$

$$\text{pmi}_{ik}^S = \log\left(\frac{p_{ik}^S}{p_{i*}^S \cdot p_{*k}^S}\right) \tag{9.7}$$

$$sc_i^S(c_k^S) = \begin{cases} \text{pmi}_{ik}^S, & \text{if pmi}_{ik}^S > 0 \\ 0, & \text{otherwise} \end{cases} \tag{9.8}$$

We may compute the similar feature score for any target language word $w_i^T$. $C(w_i^S, c_k^S)$ is the counter variable counting the frequency of the word $w_i^S \in V^S$ occurring with the context feature $c_k$ (actually $c_k^S$). $p_{ik}^S$ is the estimated probability that the word $w_i^S$ occurs with $c_k$, $p_{i*}^S$ is the estimated probability for the word $w_i^S$, while $p_{*k}^S$ is the estimated probability for the context feature $c_k$. PPMI is designed to assign a high value to $sc_i^S(c_k^S)$ when there exists an interesting semantic relation between $w_i^S$ and $c_k^S$. Otherwise, $sc_i^S(c_k^S)$ is assigned a value of zero, indicating that the relation between $w_i^S$ and $c_k^S$ is uninformative. A well-known issue with PMI (and consequently with PPMI) lies in the fact that it is biased towards infrequent events. Therefore, several smoothing schemes have been proposed to alleviate that issue. Here, we use a smoothed version of

PPMI as presented in [233, 299], where the smoothed PMI is defined as follows:

$$\gamma_{ik}^{S} = \frac{C(w_i^S, c_k^S)}{C(w_i^S, c_k^S) + 1} \cdot \frac{min(\sum_{n=1}^{N} C(w_i^S, c_n^S), \sum_{w_j^S \in V^S} C(w_j^S, c_k^S))}{min(\sum_{n=1}^{N} C(w_i^S, c_n^S), \sum_{w_j^S \in V^S} C(w_j^S, c_k^S)) + 1} \tag{9.9}$$

$$\text{smoothedpmi}_{ik}^{S} = \gamma_{ik}^{S} \cdot \text{pmi}_{ik}^{S} \tag{9.10}$$

where $pmi_{ik}^{S}$ is again computed according to eq. (9.8). $\gamma_{ik}^{S}$ denotes a smoothing weight, and *min* denotes a function which returns the minimum of two inputs.

Again, based on the results reported in the relevant literature [40, 160, 299], we opt for the cosine similarity as a standard choice for $SF$ (see eq. 6.9). We have also experimented with different window sizes ranging from 3 to 15 in both directions around the pivot word, but we have not detected any major qualitative difference in the results and their interpretation. Therefore, all reported results are obtained by setting the window size to 6.

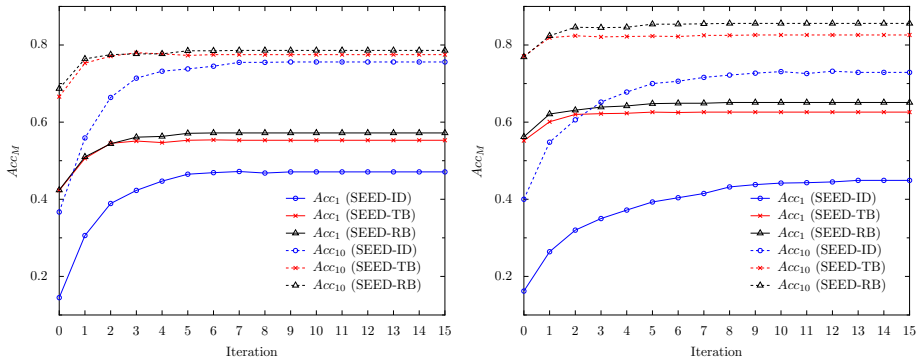
## 9.4 Experiments, Results and Discussion

### 9.4.1 Experiment I: Is Initialization Important?

In recent work, Peirsman and Padó [234, 235] report that "the size and quality of the (seed) lexicon are not of primary importance given that the bootstrapping procedure effectively helped filter out incorrect translation pairs and added more newly identified mutual nearest neighbors." According to their findings, (1) noisy translation pairs are corrected in later stages of the bootstrapping process, since the quality of cross-lingual vector spaces gradually increases, (2) the size of the seed lexicon does not matter since the bootstrapping approach is able to learn translation pairs that were previously not present in the seed lexicon. Additionally, they do not provide any insight whether the frequency of seeds in the corpus influences the quality of induced cross-lingual vector spaces. In this chapter, we question these claims with a series of BLE experiments.

All experiments conducted in this section do not rely on any extra confidence estimation except for the symmetry constraint, that is, in each step we enrich the cross-lingual vector space with all new symmetric translation pairs (see alg. 9.1 and sect. 9.2.3).

**Experimental Question 1: Same Size, Different Seeding Models?** The goal of this experiment is to test whether the quality of seeds plays an important role in the bootstrapping approach. We experiment with three different seed




(a) Spanish to English

(b) Italian to English

Figure 9.2: Results with 3 different seeding methods as starting points of the bootstrapping process: (i) identical words only (SEED-ID), (ii) the BC-Topics model (SEED-TB), (iii) the BC-Responses model (SEED-RB). (a) $Acc_1$ and $Acc_{10}$ scores for ES-EN, (b) $Acc_1$ and $Acc_{10}$ scores for IT-EN.

lexicons: (1) Following [234, 89], we harvest identically spelled words across two languages and treat them as one-to-one translations. This procedure results in 459 seed translation pairs for ES-EN, and 431 pairs for IT-EN (SEED-ID), (2) We obtain symmetric translation pairs using the BC-Topics model of similarity (see sect. 8.3 in chapter 8) and use 459 pairs that have the highest frequency in the ES-EN Wikipedia corpus as seeds for ES-EN (similarly 431 pairs for IT-EN) (SEED-TB), (3) As in (2), but we now use the BC-Responses model to acquire seeds (SEED-RB). The frequency of a one-to-one translation pair is simply computed as the geometric mean of the frequencies of words that constitute the translation pair. Fig. 9.2a and fig. 9.2b display the progress of the same bootstrapping procedure using the three different seed lexicons. We derive several interesting conclusions:

(i) *Regardless of the actual choice of the seeding method, the bootstrapping process proves its validity and utility* since we observe that the quality of induced cross-lingual vector spaces increases over time for all three seeding methods. The bootstrapping procedure converges quickly. The increase is especially prominent in the first few iterations, when the approach learns more new dimensions (see fig. 9.3).

(ii) *The seeding method is important.* A bootstrapping approach that starts with a better seed lexicon is able to extract bilingual lexicons of higher quality as reflected in $Acc_1$ scores. Although the bootstrapping approach seems more beneficial when dealing with noisier seed lexicons (226% increase in terms of

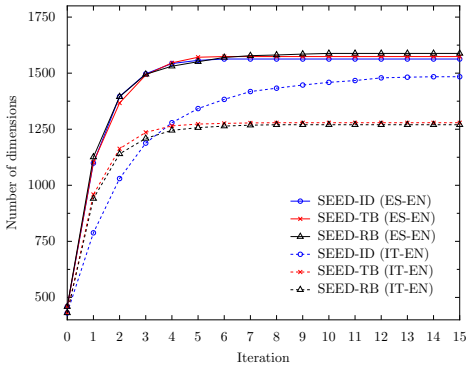


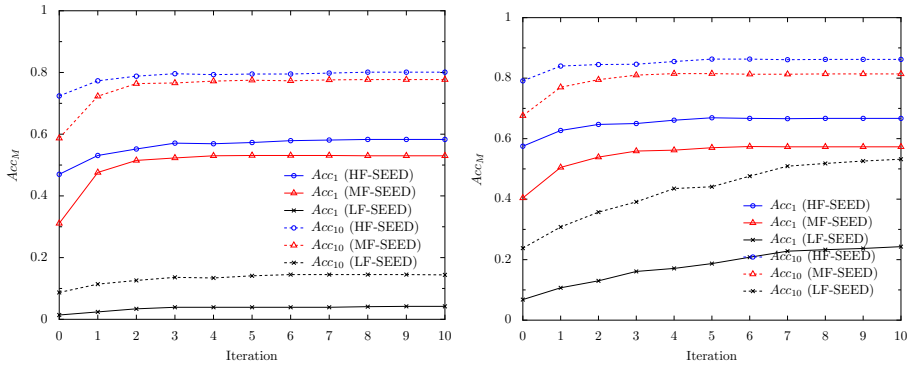

Figure 9.3: The number of dimensions in the cross-lingual vector space with the 3 different seeding methods in each iteration for ES-EN and IT-EN. The bootstrapping procedure typically converges after a few iterations.

$Acc_1$ for ES-EN and 177% increase for IT-EN when starting with SEED-ID, compared to 35% increase for ES-EN, and 15% for IT-EN with SEED-RB), when starting from a noisy seed lexicon such as SEED-ID the method is unable to reach the same level of performance. Starting with SEED-ID, the approach is able to recover noisy dimensions from an initial cross-lingual vector space, but it is still unable to match the results that are obtained when starting from a better initial space (e.g., SEED-RB).

(iii) SEED-RB produces slightly better results than SEED-TB (e.g., the final $Acc_1$ of 0.649 for SEED-RB compared to 0.626 for SEED-TB for IT-EN, and 0.572 compared to 0.553 for ES-EN). This finding is in line with the results reported in [314] and the previous chapter, where BC-Responses proved to be a more robust and a more effective method when applied to the BLE task directly. In all further experiments we use BC-Responses to acquire seed pairs, i.e., the seeding method is SEED-RB.

**Experimental Question 2: Does the Frequency of Seeds Matter?** In the next experiment, we test whether the frequency of seeds in the corpus plays an important role in the bootstrapping process. The intuition is that by using highly frequent and highly confident translation pairs the bootstrapping method has more reliable clues that help extract new dimensions in subsequent iterations. On the other hand, low-frequency pairs (although potentially correct one-to-one translations) do not occur in the corpus and in the contexts of other words frequently enough, and are therefore not sufficient to extract reliable new dimensions of the space.

To test the hypothesis, we again obtain all symmetric translation pairs using BC-Responses and then sort them in descending order based on their frequency




(a) Spanish to English (b) Italian to English

Figure 9.4: Results on the BLE task with SEED-RB when using seed translation pairs of different frequency: (i) high-frequency (HF-SEED), (ii) medium-frequency (MF-SEED), (iii) low-frequency (LF-SEED).

in the corpus. In total, we retrieve a sorted list of 2,031 symmetric translation pairs for ES-EN, and 1,689 pairs for IT-EN. Following that, we split the list in three parts of equal size: (i) the top third comprises translation pairs with the highest frequency in the corpus (HF-SEED), (ii) the middle third comprises pairs of "medium" frequency (MF-SEED), (iii) the bottom third are low-frequency pairs (LF-SEED). We then use these three different seed lexicons of equal size to seed the bootstrapping approach. Fig. 9.4a and fig. 9.4b show the progress of the bootstrapping process using these three seed lexicons. We again observe several interesting phenomena:

(i) *High-frequency seed translation pairs are better seeds*, and that finding is in line with our hypothesis. Although the bootstrapping approach again displays a positive trend regardless of the actual choice of seeds (we observe an increase even when using LF-SEED), high-frequency seeds lead to better overall results in the BLE task. Besides its high presence in contexts of other words, another advantage of high-frequency seed pairs is the fact that an initial similarity method will typically acquire more reliable translation candidates for such words [236]. For instance, 89.5% of ES-EN pairs in HF-SEED are correct one-to-one translations, compared to 65.1% in MF-SEED, and 44.3% in LF-SEED.

(ii) The difference in results between HF-SEED and MF-SEED is more visible in $Acc_1$ scores. Although both seed lexicons for all test words provide ranked lists which contain words that exhibit some semantic relation to the given word, the reliability and the frequency of translation pairs are especially important for detecting the relation of cross-lingual word synonymy, that is, the translational equivalence that is exploited in building one-to-one bilingual lexicons.

**Experimental Question 3: Does Size Matter?** The following experiment investigates whether cross-lingual vector spaces may be effectively bootstrapped from small high-quality seed lexicons, and if larger seed lexicons necessarily lead to cross-lingual vector spaces of higher quality as reflected in BLE results. We again retrieve a sorted list of symmetric translation pairs. Following that, we build seed lexicons of various sizes by retaining only the first $TOP$ pairs from the list, where we vary $TOP$ from 200 to $1,400$ in steps of 200. We also use the entire sorted list as a seed lexicon (*All*), and compare the results on the BLE task with the results obtained by applying the BC-Responses and BC-Topics similarity models directly [314]. The results are summarized in tab. 9.1 and tab. 9.2. We observe the following:

(i) If we provide a seed lexicon with sufficient entries, the bootstrapping procedure provides comparable results regardless of the seed lexicon size, although results tend to be higher for larger seed lexicons (e.g., compare results when starting with 600 and $1,200$ lexicon entries). When starting with the size of 600, the bootstrapping approach is able to find dimensions that were already in the seed lexicon of size $1,200$. The consequence is that, although bootstrapping with a smaller seed lexicon displays a slower start (see the difference in results at iteration 0), the performances level after convergence.

(ii) Regardless of the seed lexicon size, the bootstrapping approach is valuable. It consistently improves the quality of the induced cross-lingual vector space, and consequently, the quality of bilingual lexicons extracted using that vector space. The positive impact is more prominent for smaller seed lexicons, i.e., we observe an increase of 78% for ES-EN when starting with only 200 seed pairs, compared to an increase of 15% when starting with 800 seed pairs, and 10% when starting with $1,400$ seed pairs.

(iii) The bootstrapping approach outperforms BC-Responses and BC-Topics in terms of $Acc_1$ and $MRR$ scores for both language pairs when the seed lexicon provides a sufficient number of entries. However, in terms of $Acc_{10}$, BC-Topics and BC-Responses still exhibit comparable (for IT-EN) or even better (ES-EN) results. Both BC-Topics and BC-Responses are MuPTM-based methods that, due to MuPTM properties, model the similarity of two words at the level of documents as contexts, while the bootstrapping approach is a window-based approach that narrows down the context to a local neighborhood of a word. The MuPTM-based models are better suited to detect a general *topical similarity* of words, and are therefore not always able to push the real cross-lingual synonyms higher in the ranked list of semantically similar words, while the window-based bootstrapping approach is better tailored to model the relation of cross-lingual synonymy, that is, to extract one-to-one translation pairs (as reflected in $Acc_1$ scores). A similar conclusion for monolingual settings is drawn in [17].

| **Iteration:** | **0** | | | **2** | | | **5** | | | **10** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Seed lexicon** | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ |
| 200(→1617) | 0.274 | 0.352 | 0.525 | 0.446 | 0.534 | 0.713 | 0.481 | 0.569 | 0.753 | 0.488 | 0.576 | 0.752 |
| 400(→1563) | 0.416 | 0.499 | 0.663 | 0.518 | 0.602 | 0.774 | 0.542 | 0.620 | 0.787 | 0.545 | 0.625 | 0.788 |
| 600(→1554) | 0.459 | 0.539 | 0.707 | 0.550 | 0.630 | 0.787 | 0.573 | 0.650 | 0.803 | 0.578 | 0.654 | 0.802 |
| 800(→1582) | 0.494 | 0.572 | 0.728 | 0.548 | 0.631 | 0.799 | 0.563 | 0.644 | 0.802 | 0.567 | 0.646 | 0.806 |
| 1000(→1636) | 0.516 | 0.591 | 0.744 | 0.563 | 0.644 | 0.805 | 0.578 | 0.656 | 0.813 | 0.581 | 0.658 | 0.817 |
| 1200(→1740) | 0.536 | 0.613 | 0.764 | 0.586 | 0.661 | 0.804 | 0.588 | 0.664 | 0.812 | 0.591 | 0.667 | 0.814 |
| 1400(→1888) | 0.536 | 0.620 | 0.776 | 0.583 | 0.659 | 0.808 | 0.589 | 0.666 | 0.815 | 0.588 | 0.666 | 0.818 |
| All-2031(→2437) | 0.543 | 0.625 | 0.785 | 0.589 | 0.667 | 0.813 | 0.597 | 0.675 | 0.818 | **0.599** | **0.677** | 0.820 |
| *TopicBC* | 0.433 | 0.576 | 0.843 | – | – | – | – | – | – | – | – | – |
| *ResponseBC* | 0.517 | 0.635 | **0.891** | – | – | – | – | – | – | – | – | – |

Table 9.1: ES-EN: Results with different sizes of the seed lexicon. The number in the parentheses denotes the number of dimensions in the cross-lingual space after the bootstrapping procedure converges. The seeding method is SEED-RB.

| **Iteration:** | **0** | | | **2** | | | **5** | | | **10** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Seed lexicon** | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ | $Acc_1$ | $MRR$ | $Acc_{10}$ |
| 200(→1255) | 0.394 | 0.469 | 0.703 | 0.515 | 0.595 | 0.757 | 0.548 | 0.621 | 0.782 | 0.555 | 0.628 | 0.787 |
| 400(→1265) | 0.546 | 0.618 | 0.757 | 0.623 | 0.690 | 0.831 | 0.639 | 0.704 | 0.840 | 0.644 | 0.709 | 0.844 |
| 600(→1309) | 0.585 | 0.657 | 0.798 | 0.653 | 0.718 | 0.856 | 0.664 | 0.726 | 0.859 | **0.672** | 0.734 | 0.862 |
| 800(→1365) | 0.602 | 0.672 | 0.813 | 0.657 | 0.723 | 0.857 | 0.663 | 0.726 | 0.865 | 0.665 | 0.730 | 0.867 |
| 1000(→1416) | 0.616 | 0.688 | 0.828 | 0.629 | 0.706 | 0.853 | 0.636 | 0.709 | 0.857 | 0.642 | 0.714 | 0.861 |
| 1200(→1581) | 0.628 | 0.700 | 0.840 | 0.655 | 0.724 | 0.869 | 0.664 | 0.733 | 0.877 | 0.668 | **0.736** | **0.883** |
| 1400(→1749) | 0.626 | 0.701 | 0.851 | 0.654 | 0.727 | 0.867 | 0.656 | 0.728 | 0.867 | 0.661 | 0.733 | 0.874 |
| All-1689(→2008) | 0.616 | 0.695 | 0.850 | 0.643 | 0.716 | 0.860 | 0.653 | 0.724 | 0.862 | 0.654 | 0.726 | 0.866 |
| *TopicBC* | 0.578 | 0.667 | 0.834 | – | – | – | – | – | – | – | – | – |
| *ResponseBC* | 0.622 | 0.729 | 0.882 | – | – | – | – | – | – | – | – | – |

Table 9.2: IT-EN: Results with different sizes of the seed lexicon. The number in the parentheses denotes the number of dimensions in the cross-lingual space after the bootstrapping procedure converges. The seeding method is SEED-RB.

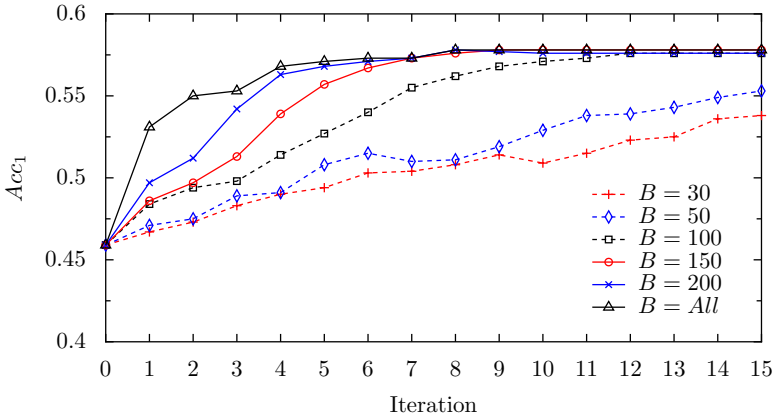



Figure 9.5: The effect of learning rate $B$ on bootstrapping. ES-EN. Seed lexicon: SEED-RB with 600 pairs, confidence function: symmetrized M-Best.

(iv) Since our bootstrapping approach utilizes BC-Responses or BC-Topics as a preprocessing step, it is obvious that the approach leads to an increased complexity. On top of the initial complexity of BC-Responses and BC-Topics, the bootstrapping method requires $|V^S||V^T|$ comparisons at each iteration, but given the fact that each $w_i^S \in V^S$ may be processed independently of any other $w_j^S \in V^S$ in each iteration, the bootstrapping method is trivially parallelizable. That makes the method computationally feasible even for vocabularies larger than the ones reported in this thesis.

### 9.4.2 Experiment II: Is Confidence Estimation Important?

According to the results from tab. 9.1 and tab. 9.2, regardless of the seed lexicon size, the bootstrapping approach does not suffer from semantic drift, i.e., if we seed the process with high-quality symmetric translation pairs, it is able to recover more pairs and add them as new dimensions of the cross-lingual vector space. However, we also study the influence of applying different confidence estimation functions on top of the symmetry constraint (see sect 9.2.3), but we do not observe any improvement in the BLE results, regardless of the actual choice of a confidence estimation function. The only observed phenomenon, as illustrated by fig. 9.5, is the *slower convergence rate* when setting the parameter $B$ to lower values.

*The symmetry constraint alone seems to be sufficient to prevent semantic drift*, but it might also be a too strong and a too conservative assumption, since only a small portion of all possible translation pairs is used to span the cross-lingual vector space (for instance, when starting with 600 entries for ES-EN, the final

cross-lingual vector space consists of only 1,554 pairs, while the total number of ES nouns is 9,439). One line of future work will address the construction of bootstrapping algorithms that also enable the usage of highly reliable asymmetric pairs as dimensions, and the confidence estimation functions might have a more important role in that setting.

## 9.5 Conclusions and Future Work

In this chapter, we have again tackled research question RQ2 and have further extended our research thread dealing with unsupervised modeling of cross-lingual semantic similarity and bilingual lexicon extraction. We have presented a new bootstrapping approach to inducing cross-lingual vector spaces from non-parallel data. The bootstrapped cross-lingual vector spaces are now spanned by bilingual lexicon entries as opposed to cross-lingual topics as in chapters 6, 7, or semantic word responses as in chapter 8) (see again sect. 6.2.2). We have again shown the utility of the induced space in the BLE task. We have systematically described, analyzed and evaluated all key components of the bootstrapping pipeline encompassing the initialization step, the updating step and the estimation of confidence for new dimensions, and convergence. Results reveal that, contrary to conclusions from prior work, the initialization of the cross-lingual vector space is especially important. We have presented a novel approach to initializing the bootstrapping procedure relying on our models of similarity described in chapters 6 and 8, and have shown that better results in the BLE task are obtained by starting from seed lexicons that comprise highly reliable high-frequent translation pairs. The bootstrapping framework presented in the paper is completely language pair independent, which makes it effectively applicable to any language pair.

In future work, one may investigate other models of similarity besides BC-Topics and BC-Responses (e.g, the method from [117]) which could be used as preliminary models for constructing an initial cross-lingual vector space. Furthermore, one may study other confidence functions and explore if asymmetric translation candidates could also contribute to the bootstrapping method. Along the same line, one may also explore whether abandoning the conservative one-to-one constraint might lead to new insights and yield more effective bootstrapping models. It is also possible to test the robustness of our fully corpus-based bootstrapping approach by porting it to more language pairs. Another interesting part of future research leads to studying how well the bootstrapping approach adapts to the corpus and whether it is possible to successfully use a similar idea of bootstrapping for *corpus transfer*, that is, learning a set of initial translation pairs from one corpus and then learning more

translation pairs on another, domain-specific corpus. It is also possible to employ the same idea of bootstrapping in the standard setting for bilingual lexicon extraction, where an external lexicon is presupposed. The bootstrapping model in that context would serve to enrich the given lexicon with more translation pairs which are specific to the particular corpus. It might lead to the better expressiveness of the cross-lingual vector space which would become better adapted to the specificities of the particular given corpus.

This chapter concludes our work on the development and design of cross-lingual semantic similarity models at the word type level. However, the following chapter continues the expansion of this complete framework for modeling semantic similarity by motivating the need for word representations at the word token level, and similarity models that perform context-sensitive computation of semantic similarity.

## 9.6 Related Publications

[1] **I. Vulić** and M.-F. Moens. "A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else)," in *Proceedings of the 2013 Conference on the Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, Washington, USA, 18-21 October 2013, pp. 1613-1624, ACL, 2013.

# 10

# Modeling Cross-Lingual Semantic Similarity in Context

*To know an object is to lead to it through a context which the world provides.*

— William James

## 10.1 Introduction

In all previous chapters we have modeled cross-lingual semantic similarity at the level of word types (see sect. 6.2.3) as it is also standard practice in the relevant literature. In short, all the models of cross-lingual semantic similarity from parallel and comparable corpora provide ranked lists of semantically similar words in the target language *in isolation* or *invariably* (i.e., the modeling is performed at the level of word types). These models do not explicitly identify and encode different senses of words (at the level of word tokens or single occurrences of word types). In practice, it means that, given a sentence *"The coach of his team was not satisfied with the game yesterday."*, these context-insensitive models of similarity are not able to detect that a Spanish word *entrenador* is more similar to a polysemous word *coach* in the context of this sentence than a Spanish word *autocar*, although *autocar* is listed as the most semantically similar word to *coach* globally/invariably without any observed context. In another example, while Spanish words *partido*, *encuentro*, *cerilla* or *correspondencia* are all highly similar to another ambiguous English word

*match* when observed in isolation, given a Spanish sentence *"She was unable to find a match in her pocket to light up a cigarette."*, it is clear that the strength of semantic similarity should change in context as only *cerilla* exhibits a strong semantic similarity to *match* within this particular sentential context.

Following this intuition, in this chapter we investigate *models of cross-lingual semantic similarity in context*. The context-sensitive models of similarity target to *re-rank* the lists of semantically similar words based on the co-occurring contexts of words. While all the previous models in our proposed MuPTM-based framework for modeling semantic similarity are context-insensitive models, in this chapter we demonstrate how to build context-aware models of semantic similarity within the same framework which relies on the same shared set of latent cross-lingual topics/concepts.

Since the work in this chapter pioneers the construction of context-aware models of cross-lingual similarity and reports on ongoing research, the presentation in this chapter will focus on the *probabilistic models of similarity* which operate with conditional topic distributions (see sect. 6.3.1 in chapter 6). In other words, we present *a probabilistic framework* for modeling context-aware models of similarity, which relies on the set of latent cross-lingual topics/concepts induced from non-parallel data.

Within this probabilistic framework, each word, regardless of its actual language, is observed and represented as *a distribution over the set of latent cross-lingual topics/concepts*. The set of latent cross-lingual topics may be observed as a set of *latent senses* hidden within data. A change in word meaning after observing its context is reflected in a change of its distribution over the latent topics. In other words, the change in meaning is accomplished as a *modulation* of the original *a priori* out-of-context distribution. In practice, it means that both out-of-context and contextualized word representations are provided in the same shared cross-lingual latent space spanned by the latent cross-lingual topics/concepts. In summary, the main contributions of this chapter are:

(i) We present a new approach to modeling cross-lingual semantic similarity in context within the same probabilistic framework described in chapter 6, which is based on the latent cross-lingual topics/concepts induced from non-parallel data. Note that the work reported in this chapter is the first research on context-aware models of semantic similarity in the cross-lingual setting.

(ii) We propose and evaluate a method of context sorting and pruning that reduces contextual noise and captures only the most informative context features. The pruning method operates in the same latent cross-lingual semantic space spanned by the latent topics.

(iii) We evaluate the proposed models, and provide results on two evaluation

datasets and three language pairs for the task of *word translation in context*. The results clearly demonstrate the utility of the context-sensitive models of cross-lingual semantic similarity, since the "contextualized" models of similarity significantly outscore context-insensitive models in this task.

The chapter is structured as follows. After reviewing related work in sect. 10.2, in sect. 10.3 we motivate the usage of latent cross-lingual topics in modeling cross-lingual similarity in context and propose a method for context sorting and pruning that is also supported by the same set of latent topics. In sect. 10.4, we describe our probabilistic approach to modeling word cross-lingual semantic similarity in context, and introduce three new context-aware similarity models. Our experimental setup and evaluation procedure are discussed in sect. 10.5, while sect. 10.6 displays results obtained for three language pairs on two different datasets accompanied by a thorough discussion. Conclusions and future work are summarized in sect. 10.7.

## 10.2 Related Work

The natural shift (or rather extension) of interest from context-insensitive models towards context-aware models of semantic similarity has already occurred in the monolingual setting, and there has been a large body of work recently that has been dedicated to modeling monolingual semantic similarity in context. However, the extension of these context-sensitive models of similarity to operate with multilingual data and to measure semantic similarity across languages is an understudied problem, and has been overlooked in the relevant literature. Therefore, in this chapter we tackle the more difficult cross-lingual setting and present a unified general probabilistic framework for context-aware modeling that does not change its modeling premises regardless of the actual setting (monolingual vs. cross-lingual or multilingual).

In the monolingual setting, several families of context-aware models have been established over the years. One line of relevant research aims to overcome the problem with distributional vector space models of similarity which compute a single word type vector for each word type by simply summing up over all occurrences of the same word type. The context-aware models that originated from these vector space models tackle that issue by introducing an idea of combining the word vectors of a source word, target word and words occurring in their contexts in order to provide a disambiguated vector representation. This idea has been elaborated in [206, 84, 286, 287], where the main difference lies in the choice of input vector representation and in the vector combination function they employ (e.g., addition or point-wise multiplication of vector elements).

However, it has been shown [78] that all these models are essentially equivalent and produce comparable results once syntactic information is ignored. A slightly different approach has been taken by [85, 251, 248]. In short, they introduce a multi-prototype vector space model that builds a different word vector for each possible meaning of a word obtained by an unsupervised clustering algorithm (e.g., they move from type- to token-based representations). Following that, they select a set of token vectors which are similar to the current context and employ only these to obtain a disambiguated representation by combination, or compute the similarity by simply averaging over all pairs of prototypes (see, e.g., [251]). Additionally, since vector composition models construct representations that go beyond individual words (e.g., representations of phrases or sentences in the same space), they also obtain word meanings in context by default. Therefore, it is also worth to mention related work on compositionality in semantic spaces, although this work does not tackle the problem of word meaning in context explicitly [272, 263, 18, 207, 110, 274, 275, 25, 50, 146, 142]. In summary, these models still need a further generalization to the multilingual/cross-lingual setting where it is unclear how to bridge the gap between two languages while retaining the key modeling premises which constitute the core of all these models.

The focus of the research reported in this chapter lies on probabilistic models of semantic similarity that rely on the induction of some sort of latent structure from data, typically in a form of *latent senses* or *latent concepts* (see sect. 10.1 and sect. 4.3 in chapter 4) [255, 227, 76, 77, 303, 228]. In the monolingual setting, the probabilistic latent variable-based models have proven to be superior to the aforementioned vector space models and algebraic models in tasks such as selectional preferences [255, 227], word similarity in context [76, 228] or lexical substitution [76, 228]. The work reported in this chapter may be observed as a generalization of this probabilistic framework that relies on the induction of latent senses/concepts to the cross-lingual setting (e.g., the work from Dinu and Lapata [76, 77] is subsumed within our framework). It will allow for future experimentations and construction of more elaborate models of similarity in context that capture more evidence (e.g., syntactic dependencies) within the same framework.

Another line of related research tackles more specifically the tasks of cross-lingual lexical substitution [200] and cross-lingual word sense disambiguation [169, 170]. Unlike our framework presented in this chapter, these models of cross-lingual lexical substitution and word sense disambiguation do not tackle nor discuss the more general notion of cross-lingual semantic similarity, since they are completely task-driven and rely on a variety of external resources which help them accomplish the specific tasks. These models typically rely at least on word alignments extracted from sentence-aligned parallel data or on end-to-end SMT systems [270, 8, 9, 262, 305, 42], or additional dictionaries, thesauri and

translation resources such as Google Translate, Babylon Dictionary, Spanishdict [19, 12, 325, 115], WordNet [184] or encyclopedic Wikipedia knowledge [19]. Our presentation in this chapter is more general and goes beyond the specificities of these tasks. Moreover, to the best of our knowledge, the work presented in this chapter is the first completely corpus-based method that relies on non-parallel data for modeling cross-lingual semantic similarity in context.

## 10.3 Towards Context-Sensitive Models of Cross-Lingual Semantic Similarity

Recall from chapter 6 that each word, regardless of its actual language, may be represented as a feature vector in a $K$-dimensional latent semantic space spanned by latent cross-lingual topics (see sect. 6.3.1). In case when these features are conditional topic probability scores $P(z_k|w_1^S)$, each word, irrespective to the language, is actually represented as a distribution over the $K$ latent topics. The $K$-dimensional vector representation of a word $w_1^S \in V^S$ is:

$$vec(w_1^S) = [P(z_1|w_1^S), P(z_2|w_1^S), \ldots, P(z_K|w_1^S)] \tag{10.1}$$

We may also represent any target language word $w_2^T$ in the same shared cross-lingual semantic space by a $K$-dimensional vector with scores $P(z_k|w_2^T)$, $k = 1, \ldots, K$, computed in the exact same fashion, and compute the similarity between $w_1^S$ and $w_2^T$ by employing a similarity function on these vector representations in the semantic space which results in a series of similarity models discussed in chapter 6. All these context-insensitive models relying on the latent cross-lingual semantic space spanned by latent cross-lingual topics use only global co-occurrence statistics from the training set and do not take into account any contextual information. They provide only *out-of-context word representations* at the word type level (see sect. 6.2.3) and are therefore able to deliver only *lists of semantically similar words in isolation.* This insight motivates us to move towards context-sensitive models of cross-lingual similarity.

A quick note on terminology throughout this chapter: the reader must be aware that the latent topics/concepts/senses are not by any means lexicographic categories of meaning, and they rather denote coarse-grained senses (or, more generally, soft clusters) induced directly from data. In addition, these latent "senses" are not word-specific nor language-specific, but global (i.e., shared across all words in both languages).

### 10.3.1 Why Context-Sensitive Models of Cross-Lingual Semantic Similarity?

The probabilistic framework relying on latent cross-lingual topics, due to its modeling properties, provides an implicit word sense disambiguation, but that disambiguation feature has not been previously exploited in related work on cross-lingual semantic word similarity. If some word typically exhibits more than one meaning, it means that the word is often related to more than one latent semantic concept. Therefore, it will be strongly associated with two or more corresponding latent topics that describe these concepts. For instance, in the monolingual setting, given the English word *plant*, we expect that word to be strongly associated with a latent concept/topic related to *Energy*, which is in English represented by high probabilities over words such as *power, industry, production, generator, powerhouse*, etc. We also expect that word to be strongly associated with a latent concept/topic related to *Biology* with words *organism, seed, green, chlorophyll*, etc. When obtaining a list of semantically similar words without being aware of the surrounding context, words from both topics will be almost equally represented in the list. The same reasoning is valid in the cross-lingual setting (see fig. 10.1).

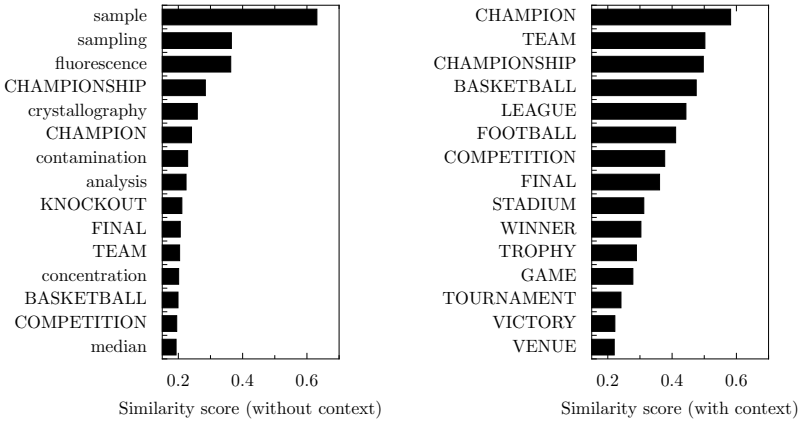



Figure 10.1: An example of cross-lingual word similarity without and with context. The lists contain English words similar to Italian word *campione* before observing any context and after observing a context word *squadra (team)*.

Here, we have obtained a list of English words similar to Italian word *campione*, which can mean *sample* or *champion* depending on its context. The similarity function is the Bhattacharyya coefficient and the model is DIRECT-FUSION (see later in sect. 10.4). English words that are semantically similar to the

English translation *sample* are given in small letters, while English words that are semantically similar to *champion* are given in capital letters.

We observe that, without any context, words that are related to both meanings of *campione* are listed as its semantically similar words. Without any context, the probability scores $P(z_k|campione)$ will be high both for concepts/topics related to science (where it is translated as *sample*), and concepts/topics related to sports (where translated as *champion*). However, the context-insensitive models of cross-lingual semantic similarity are always linking the globally dominant meaning as the most similar word to *campione*. The dominant meaning relies on the co-occurrence counts from a training corpus. In this case, without taking any context into account, each occurrence of *campione* will be strongly associated with the English word *sample* even when its context includes words such as *squadra (team)* or *calciatore (soccer player)* that come as strong indicators that *campione* is more similar to *champion* than to *sample* within this context. Therefore, our goal is to design models of cross-lingual semantic similarity that are token-based rather than only type-based. These models should be able to provide ranked lists of semantically similar words taking into account both the observed source word and its local context. The models should rely on the global estimates from a large corpus (i.e., the semantic similarity at the word type level) in cases when the local context alone (i.e., the similarity at the word token level) is not informative enough to obtain semantically similar words for the given word occurrence. For instance, take the following example sentence in Italian: *"Abbiamo bisogno di un altro campione."*. The translation of this sentence in English is *"We need another sample/champion."*. However, it is visible that we cannot deduce the correct translation from the sentential context.[1] In that case, there are two potential solutions: (S1) Use the *dominant meaning heuristic*, that is, in case where no useful contextual information is available, steer the context-sensitive models of similarity towards dominant meaning, (S2) If possible, broaden the context scope beyond the sentence limits to find extra contextual clues within the discourse of the sentence. In this chapter, following similar work in the monolingual setting [131] we investigate models that tackle S1, while we leave S2 for future work. Note that S2 is possible only if a sentence is provided within its discourse, while S1 is tailored to model the context-sensitive semantic similarity within the boundaries of an isolated sentence.

[1] As an illustration, we have also tried to translate the sentence *"We need another champion."* by the *Google Translate* tool from English to Italian and then back to English, and the obtained translation chain is as follows: *"We need another champion."* → *"Abbiamo bisogno di un altro campione."* → *"We need another sample."*

### 10.3.2 Defining Context

The context $Con(w_1^S)$ may include all words that occur in the same document or paragraph with the particular occurrence of $w_1^S$ (i.e., a *document-based context*) [159], all words occurring in the same sentence [189, 131] (i.e., a *sentence-based* or *sentential context*), only neighboring words occurring in a window of predefined size (i.e., a *window-based context*) [247, 102, 40], or only neighboring words with a particular syntactic relation to $w_1^S$ (i.e., a *dependency-based context*) [177, 231]. In this chapter we do not investigate the influence of context granularity and context type. Following the recent work from Huang et al. [131] in the monolingual setting, we limit the contextual scope to the *sentential context*. However, we emphasize that the proposed models are designed to be fully functional regardless of the actual chosen context granularity. Even more importantly, the whole global discourse is captured through latent cross-lingual concepts and their conditional topic probability scores.

Given an occurrence of a word $w_1^S$, we build its context set $Con(w_1^S) = \{cw_1^S, \ldots, cw_r^S\}$ that comprises $r$ words from $V^S$ that co-occur with $w_1^S$ in a defined *contextual scope*, e.g., when operating in the sentential context, $Con(w_1^S)$ consists of words occurring in the same sentence with the particular instance of $w_1^S$. Following Mitchell and Lapata [206], for the sake of simplicity, we impose the *bag-of-words* assumption, and do not take into account the order of words in the context set as well as context words' dependency relations to $w_1^S$. Investigating different context types is a subject of future work.

**Context Sorting and Pruning.** By using all words occurring with $w_1^S$ in a context set (e.g., a sentence) to build the set $Con(w_1^S)$, we do not make any distinction between "informative and "uninformative" context words. However, some context words bear more contextual information about the observed word $w_1^S$ and are stronger indicators of the correct word meaning in that particular context. For instance, in the sentence *"The coach of his team was not satisfied with the game yesterday"*, words *game* and *team* are strong clues that *coach* should be translated as *entrenador* while the context word *yesterday* does not bring any extra contextual information that could resolve the ambiguity in meaning of *coach*.

Therefore, in the final context set $Con(w_1^S)$ it is useful to retain only the context words that really bring extra semantic information. We achieve that by exploiting the same latent semantic space to provide the similarity score between the observed word $w_1^S$ and each word $cw_i^S, i = 1, \ldots, r$ from its context set $Con(w_1^S)$. Each word $cw_i^S$ may be represented by its vector $vec(cw_i^S)$ (see eq. (10.1)) in the same latent semantic space, and there we can compute the similarity between its vector and $vec(w_1^S)$. We can then sort the similarity

scores for each $cw_i^S$ and retain only the top scoring $M$ context words in the final set $Con(w_1^S)$. The procedure of context sorting and pruning should improve the semantic cohesion between $w_1^S$ and its context since only informative contextual words/features are now present in $Con(w_1^S)$, and we reduce the noise coming from uninformative contextual features that are not semantically related to $w_1^S$.

### 10.3.3 Projecting Context into the Latent Semantic Space

The probabilistic framework that is supported by latent cross-lingual topics allows for having the $K$-dimensional vector representations in the same latent semantic space spanned by cross-lingual topics for: (1) Single words regardless of their actual language, and (2) Sets that comprise multiple words. Therefore, *we are able to project the observed source word, all target words, and the context set of the observed source word to the same latent semantic space spanned by latent cross-lingual topics.*

Eq. (10.1) shows how to represent single words in the latent semantic space. Now, we present a way to address compositionality, that is, we show how to build the same representations in the same latent semantic space beyond the word level. We need to compute a conditional topic distribution for the context set $Con(w_1^S)$, that is, we have to compute the probability scores $P(z_k|Con(w_1^S))$ for each $z_k \in \mathcal{Z}$. Remember that the context $Con(w_1^S)$ is actually a set of $r$ (or $M$ after pruning) words $Con(w_1^S) = \{cw_1^S, \ldots, cw_r^S\}$ (see sect. 10.3.2). Under the *single-topic assumption* [114] and following Bayes' rule, it holds:

$$P(z_k|Con(w_1^S)) = \frac{P(Con(w_1^S)|z_k)P(z_k)}{P(Con(w_1^S))} = \frac{P(cw_1^S, \ldots, cw_r^S|z_k)P(z_k)}{\sum_{l=1}^{K} P(cw_1^S, \ldots, cw_r^S|z_l)P(z_l)} \tag{10.2}$$

$$= \frac{\prod_{j=1}^{r} P(cw_j^S|z_k)P(z_k)}{\sum_{l=1}^{K} \prod_{j=1}^{r} P(cw_j^S|z_l)P(z_l)} \tag{10.3}$$

Note that here we use a simplification where we assume that all $cw_j^S \in Con(w_1^S)$ are conditionally independent given $z_k$. The assumption of the conditional independence of unigrams is a standard heuristic applied in *bag-of-words* models in NLP and IR (e.g., observe a direct analogy to probabilistic language models for IR where the assumption of independence of query words is imposed [242, 126, 163]), but we have to forewarn the reader that in general the equation $P(cw_1^S, \ldots, cw_r^S|z_k) = \prod_{j=1}^{r} P(cw_j^S|z_k)$ is not exact. However, by adopting the conditional independence assumption, in case of the uniform topic prior $P(z_k)$ (i.e., we assume that we do not possess any prior knowledge about the

importance of latent topics in a multilingual corpus), eq. (10.3) may be further simplified:

$$P(z_k|Con(w_1^S)) \approx \frac{\prod_{j=1}^{r} P(cw_j^S|z_k)}{\sum_{l=1}^{K}\prod_{j=1}^{r} P(cw_j^S|z_l)} \tag{10.4}$$

The representation of the context set in the latent semantic space is then:

$$vec(Con(w_1^S)) = [P(z_1|Con(w_1^S)), \ldots, P(z_K|Con(w_1^S))] \tag{10.5}$$

We can then compute the similarity between words and sets of words given in the same shared latent semantic space in a uniform way, irrespective of their actual language.[2] We are also able to compute the scores $P(w_2^T|Con(w_1^S))$ in the same fashion as $P(w_2^T|w_1^S)$ (e.g., for the Cue model from sect. 6.3.5). Namely, $P(w_2^T|Con(w_1^S))$ is simply computed as:

$$P(w_2^T|Con(w_1^S)) = \sum_{k=1}^{K} P(w_2^T|z_k)P(z_k|Con(w_1^S)) \tag{10.6}$$

where $P(z_k|Con(w_1^S))$ is given by eq. (10.4), and $P(w_2^T|z_k)$ obtained from a multilingual topic model directly. We use all these properties when building our context-sensitive models of cross-lingual similarity.

## 10.4 Context-Sensitive Models of Similarity via Latent Cross-Lingual Topics

The models of cross-lingual semantic similarity in context described in this section rely on the representations of words and their context sets in the same latent semantic space spanned by latent cross-lingual concepts/topics as discussed in sect. 10.3. The models differ in the way the contextual knowledge is fused with the isolated out-of-context word representations.

The key idea behind these models is to represent a word $w_1^S$ in the latent semantic space as a distribution over the latent cross-lingual topics, but now

[2] An additional remark on the context set: Words in the set do not necessarily have to be in the same language as word $w_1^S$. Imagine a scenario where a translator translates a sentence from the source language $L_S$ to the target language $L_T$, and she/he wants to know the correct translation of some polysemous word in $L_S$. The translator can use words in $L_S$ that surround that word as context words, but she can also enrich the set of context words with content words from $L_T$ that have already been translated in the process. Investigating different ways of building the context set is beyond the scope of this work, but we want to stress that the modeling principles remain the same regardless of the chosen context set.

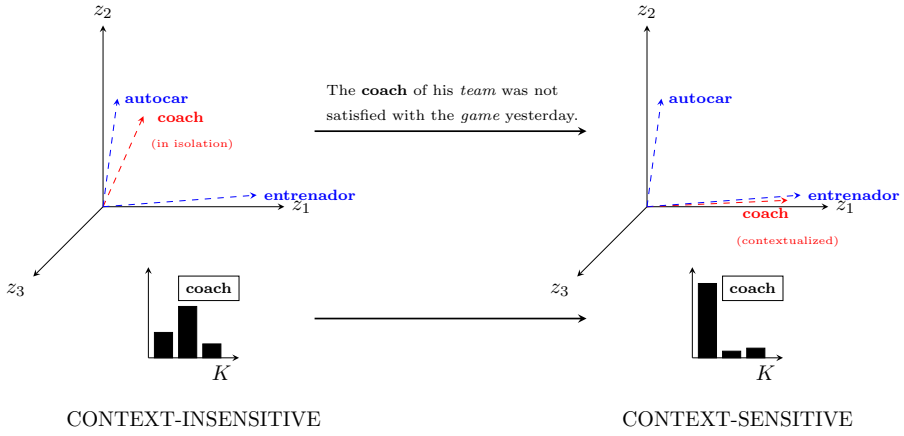



Figure 10.2: An illustrative toy example of the main intuitions in our probabilistic framework for building context-sensitive models of cross-lingual semantic similarity.

with an additional modulation of the representation after taking its local context into account. The modulated word representation in the semantic space spanned by $K$ latent cross-lingual concepts is then:

$$vec(w_1^S, Con(w_1^S)) = [P'(z_1|w_1^S), \ldots, P'(z_K|w_1^S)] \tag{10.7}$$

where $P'(z_K|w_1^S)$ denotes the recalculated (or modulated) probability score for the conditional concept/topic distribution of $w_1^S$ after observing its context $Con(w_1^S)$. For an illustration of the key idea, see fig. 10.2. The figure shows a cross-lingual semantic space spanned by only three latent cross-lingual topics (axes $z_1$, $z_2$ and $z_3$): A change in meaning is reflected as a change in a probability distribution over latent cross-lingual topics that span a shared latent semantic space. A change in the probability distribution may then actually steer an English word *coach* towards its correct (Spanish) meaning in context: Although the word when given in isolation is more similar to a Spanish word *autocar*, the similarity scores (i.e., in this case the distances in the latent semantic spaces) after observing the sentential context of the word are recalculated according to the context, and *coach* is now more similar to another Spanish word *entrenador*.

The intuition is that the context helps to disambiguate the true meaning of the occurrence of the word $w_1^S$. In other words, after observing the context of the word $w_1^S$, fewer latent cross-lingual topics will share most of the probability mass in the modulated context-aware word representation.

### 10.4.1 DIRECT-FUSION Model

The first approach makes the conditional distribution over latent cross-lingual topics from eq. (10.7) directly dependent on both word $w_1^S$ and its context $Con(w_1^S)$. The probability score $P'(z_k|w_1^S)$ from eq. (10.7) for each $z_k \in \mathcal{Z}$ is then given as $P'(z_k|w_1^S) = P(z_k|w_1^S, Con(w_1^S))$.

Going back to the example from fig. 10.1, after observing an extra context word *squadra (team)*, the conditioning of the topics, that is, the conditional topic probability scores $P(z_k|campione, Con(campione))$ are recalculated, where $Con(campione) = \{squadra\}$ now denotes the context of the word *campione*. Due to the observed context, more probability mass is now assigned to the concepts/topics related to sports. Because of the constraint $\sum_{k=1}^{K} P(z_k|campione, Con(campione)) = 1$, it automatically means that the importance of topics related to science decreases. Therefore, only words related to sports are now present in the final list of the most semantically similar words to this occurrence of word *campione* (see fig. 10.1) and the most semantically similar word to the Italian word *campione* is now *champion* instead of *sample*.

We still have to estimate the probability $P(z_k|w_1^S, Con(w_1^S))$, that is, the probability that word $w_1^S$ is assigned to the latent concept/topic $z_k$ given its context $Con(w_1^S)$. We can write:

$$P(z_k|w_1^S, Con(w_1^S)) = \frac{P(z_k, w_1^S)P(Con(w_1^S)|z_k)}{\sum_{l=1}^{K} P(z_l, w_1^S)P(Con(w_1^S)|z_l)} \tag{10.8}$$

Since $P(z_k, w_1^S) = P(w_1^S|z_k)P(z_k)$, if we closely follow the derivation from eq. (10.3) which shows how to project a set of words into the latent semantic space (and again assume the uniform prior $P(z_k)$), we obtain the following formula:

$$P'(z_k|w_1^S) = P(z_k|w_1^S, Con(w_1^S)) = \frac{P(w_1^S|z_k)\prod_{j=1}^{r} P(cw_j^S|z_k)}{\sum_{l=1}^{K} P(w_1^S|z_l)\prod_{j=1}^{r} P(cw_j^S|z_l)} \tag{10.9}$$

The ranking of all words $w_2^T \in V^T$ according to their similarity to $w_1^S$ may be computed by detecting the similarity score between their representation in the $K$-dimensional latent semantic space and the modulated source word representation as given by eq. (10.7) and eq. (10.9) again using any of the existing similarity functions [167, 43]. The similarity score $sim(w_1^S, w_2^T, Con(w_1^S))$ between some $w_2^T \in V^T$ represented by its vector $vec(w_2^T)$ and the observed word $w_1^S$ given its context $Con(w_1^S)$ is computed as:

$$sim(w_1^S, w_2^T, Con(w_1^S)) = SF\Big(vec\big(w_1^S, Con(w_1^S)\big), vec\big(w_2^T\big)\Big) \tag{10.10}$$

Words are then ranked according to their respective similarity scores and the best scoring candidate may be selected as the translation candidate for an occurrence of the word $w_1^S$ given its local context.

Since the contextual knowledge is integrated directly into the estimation of probability $P(z_k|w_1^S, Con(w_1^S))$, we name this context-aware model of cross-lingual semantic similarity the *DIRECT-FUSION* model.

### 10.4.2 SMOOTHED-FUSION Model

The next model follows the modeling paradigm established within the framework of language modeling (LM), where the idea is to "back off" to a lower order $N$-gram in case we do not possess any evidence about a higher-order $N$-gram [141]. A similar approach based on smoothing is utilized in probabilistic parsing [51, 151, 52] and POS tagging [324, 36]. An identical idea has also been exploited in the LM framework for information retrieval [242, 53]. The idea behind the IR LM approach is to smooth the probability of a word in a document by a probability that the same word will occur in the entire document collection. We adopt a similar principle in this chapter, where the idea now is to smooth the representation of a word in the latent semantic space induced only by the words in its local context with the out-of-context type-based representation of that word induced directly from a large training corpus. In other words, the modulated probability score $P'(z_k|w_1^S)$ from eq. (10.7) is calculated as:

$$P'(z_k|w_1^S) = \lambda_1 P(z_k|Con(w_1^S)) + (1 - \lambda_1) P(z_k|w_1^S) \tag{10.11}$$

where $\lambda_1$ is the interpolation parameter, $P(z_k|w_1^S)$ is the out-of-context conditional topic probability score as in eq. (10.1), and $P(z_k|Con(w_1^S))$ is given by eq. (10.3).

This model compromises between the pure contextual word representation and the out-of-context word representation. In cases when the local context of word $w_1^S$ is informative enough, the factor $P(z_k|Con(w_1^S))$ is sufficient to provide the ranking of terms in $V^T$, that is, to detect words that are semantically similar to $w_1^S$ based on its context. However, if the context is not reliable, we have to smooth the pure context-based representation with the out-of-context word representation (the factor $P(z_k|w_1^S)$). We call this model the *SMOOTHED-FUSION* model.

The ranking of words $w_2^T \in V^T$ then finally proceeds in the same manner as in the DIRECT-FUSION model following eq. (10.10), but now using eq. (10.11) for the modulated probability scores $P'(z_k|w_1^S)$.

### 10.4.3 LATE-FUSION Model

The last model is conceptually similar to the SMOOTHED-FUSION model, but it performs smoothing at a later stage. It proceeds in two steps: (1) Given a target word $w_2^T \in V^T$, the model computes similarity scores separately between (i) the context set $Con(w_1^S)$ and $w_2^T$, and (ii) the word $w_1^S$ in isolation and $w_2^T$ (again, at the word type level); (2) It linearly combines the obtained similarity scores. More formally, we may write:

$$sim(w_1^S, w_2^T, Con(w_1^S)) = \lambda_2 SF\Big(vec\big(Con(w_1^S)\big), vec\big(w_2^T\big)\Big) + (1-\lambda_2) SF\Big(vec\big(w_1^S\big), vec\big(w_2^T\big)\Big) \tag{10.12}$$

where $\lambda_2$ is the interpolation parameter. Since this model computes the similarity with each target word separately for the source word in isolation and its local context, and combines the obtained similarity scores after the computations, this model is called *LATE-FUSION*.

## 10.5 Experimental Setup

In this section, we first describe the evaluation task of word translation in context in which we demonstrate the utility of our context-sensitive models of cross-lingual semantic similarity. Following that, we provide an overview of our experimental setup, with an emphasis on a new resource for testing which, unlike all previous related test datasets, builds a small repository of ambiguous words in languages other than English (e.g., Italian, Spanish and Dutch), and allows for testing the models of similarity in the $X \rightarrow English$ direction (while all previous test sets provided only the $English \rightarrow X$ direction, e.g., [169, 168, 200, 169]).

### 10.5.1 Evaluation Task: Word Translation in Context

Given an occurrence of a polysemous word $w_1^S \in V^S$ in the source language $L_S$ with vocabulary $V^S$, the task is to choose the correct translation in the target language $L_T$ of that particular occurrence of $w_1^S$ from the given set $\mathcal{TC}(w_1^S) = \{t_1^T, \ldots, t_{tq}^T\}$, $\mathcal{TC}(w_1^S) \subseteq V^T$, of its $tq$ possible translations/meanings. We may refer to $\mathcal{TC}(w_1^S)$ as an inventory of translation candidates for $w_1^S$. The task of *word translation in context* may be interpreted as ranking the $tq$ translation candidates with respect to the observed local context $Con(w_1^S)$ of the occurrence of the word $w_1^S$. The best scoring translation candidate in the

ranked list is then the correct translation for that particular occurrence of $w_1^S$ observing its local context $Con(w_1^S)$.

In that respect, the task is very similar to the *lexical substitution task* in monolingual settings [192] and across languages [200]. The difference here is that the task is not to compute the plausibility of each potential cross-lingual lexical substitute, but to propose a single most likely translation given a word and its context. Moreover, since each $\mathcal{TC}$ may be observed as a cross-lingual sense inventory, the task is almost equivalent to the task of *cross-lingual word sense disambiguation* [169, 170], where the task is to propose the correct sense or a set of correct senses in the target language given a polysemous word in the source language and its sense inventory.

### 10.5.2 Training, Testing and Evaluation

**Training Collections.** We use exactly the same training collections for three language pairs (Spanish-English (ES-EN), Italian-English (IT-EN), and Dutch-English (NL-EN) trained on Wiki and Wiki+EP) as in chapter 8, and employ exactly the same preprocessing steps. The BiLDA model is trained on these collections with Gibbs sampling with the standard parameter setting again as in chapters 7-9: $K = 2000$, $\alpha = 50/K$, $\beta = 0.01$ [278, 314, 315].

**Test Datasets.** We use two different evaluation datasets.

**Test Dataset I: CWT+JA-BNC**. The first dataset is the benchmarking evaluation dataset for the task of cross-lingual word sense disambiguation [169]. This evaluation dataset consists of 20 English (EN) polysemous nouns and, for each noun, it contains 50 EN sentences in which the noun occurs (hence there are $1,000$ test sentences in total), taken from the JRC-ACQUIS corpus [276] and the British National Corpus (BNC). The sense inventory $\mathcal{TC}$ for these nouns was created from Europarl. The complete dataset and sense inventory construction procedure are presented by Lefever and Hoste [168] and we refer the interested reader to check the details in their paper. In short, different instances of the same noun type in the sentential context capture different meanings of the noun, and human annotators were asked to check and rank potential translation candidates (i.e., potential senses) in five other languages for each occurrence of each EN noun in the sentential context. In this thesis we use Italian (IT), Spanish (ES) and Dutch (NL) as target languages. We build the list of translation candidates $\mathcal{TC}$ for each English noun by harvesting all possible translations of that noun as given in the sense inventory. This is our *CWT+JA-BNC* evaluation dataset. The dataset also includes additional 5 polysemous EN nouns with 20 sentences each as a development set. This subset of 100 sentences is utilized to tune the parameters of our models. We

| **Spanish** | **Italian** | **Dutch** |
|---|---|---|
| *Ambiguous word* (Possible senses/translations) | *Ambiguous word* (Possible senses/translations) | *Ambiguous word* (Possible senses/translations) |
| *1. estación* (station; season) | *1. raggio* (ray; radius; spoke) | *1. toren* (rook; tower) |
| *2. ensayo* (essay; rehearsal; trial) | *2. accordo* (chord; agreement) | *2. beeld* (image; statue) |
| *3. núcleo* (core; kernel; nucleus) | *3. moto* (motion; motorcycle) | 3. *blade* (blade; leaf; magazine) |
| *4. vela* (sail; candle) | *4. calcio* (calcium; football; stock) | *4.fusie* (fusion; merger) |
| *5. escudo* (escudo; escutcheon; shield) | *5. terra* (earth; land) | *5. stam* (stem; trunk; tribe) |
| *6. papa* (Pope; potato) | *6. tavola* (board; panel; table) | *6. koper* (copper; buyer) |
| *7. cola* (glue; coke; tail; queue) | *7. campione* (champion; sample) | *7. bloem* (flower; flour) |
| *8. cometa* (comet; kite) | *8. carta* (card; paper; map) | *8. spanning* (voltage; tension; stress) |
| *9. disco* (disco; discus; disk) | *9. piano* (floor; plane; plan; piano) | *9. noot* (note; nut) |
| *10. banda* (band; gang; strip) | *10. disco* (disco; discus; disk) | *10. akkoord* (chord; agreement) |
| *11. cinta* (ribbon; tape) | *11. istruzione* (education; instruction) | *11. munt* (coin; currency; mint) |
| *12. banco* (bank; bench; shoal) | *12. gabinetto* (cabinet; office; toilet) | *12. pool* (pole; pool) |
| *13. frente* (forehead; front) | *13. torre* (rook; tower) | *band* (band; tyre; tape) |
| *14. fuga* (escape; fugue; leak) | *14. campo* (camp; field) | *14. kern* (core; kernel; nucleus) |
| *15. gota* (gout; drop) | *15. gomma* (rubber; gum; tyre) | *15. kop* (cup; head) |

Table 10.1: Sets of 15 ambiguous words in Spanish, Italian and Dutch from the CWT+Wiki dataset accompanied by the sets of their respective possible senses/translations in English.

use the CWT+JA-BNC test dataset to evaluate our context-sensitive models of cross-lingual similarity in the task of EN-to-ES/IT/NL word translation in context.

**Test Dataset II: CWT+Wiki**. We have constructed another dataset in a similar fashion, but in the opposite translation direction, now dealing with polysemous words in Spanish, Italian and Dutch and aiming to find their correct translation in English given the sentential context. We have selected 15 polysemous nouns and have manually extracted 24 sentences for each word that capture different meanings of the noun from Wikipedia (as an example,

| **Sentence in English** (CWT+JA-BNC)<br>**(Correct Translation (IT))** |
|---|
| 1. I'll buy a train or **coach** ticket.<br>**(autobus)**<br>2. In fact, the **coach** - drawn by two grey horses - was only called into service...<br>**(dilligenza)**<br>3. If any team member is at all suspect, the **coach** should put them into third position.<br>**(allenatore)**<br>4. On the international occasional carriage of passengers by **coach** and bus.<br>**(treno)** |
| **Sentence in Italian** (CWT+Wiki)<br>**(Correct Translation (EN))** |
| 1. I primi **calci** furono prodotti in legno ma recentemente...<br>**(stock)**<br>2. In caso di osteoporosi si verifica un eccesso di rilascio di **calcio** dallo scheletro...<br>**(calcium)**<br>3. La crescita del **calcio** femminile professionistico ha visto il lancio di competizioni...<br>**(football)**<br>4. Il **calcio** di questa pistola (Beretta Modello 21a, calibro .25) ha le guancette...<br>**(stock)** |

Table 10.2: Example sentences from our CWT+JA-BNC and CWT+Wiki evaluation datasets with the corresponding correct word translations from the ground truth.

see tab. 10.1 for an overview of polysemous words in Spanish, Italian, and Dutch along with the set of their possible meanings in English). Since we noticed an imbalance in CWT+JA-BNC, that is, for some nouns the single dominant meaning is considered correct for almost all 50 instances of that word in the CWT+JA-BNC dataset, here we have decided to design a more balanced evaluation dataset. In case of $tq$ different translation candidates in $\mathcal{TC}(w_1^S)$ for some word $w_1^S$, the dataset contains exactly $24/tq$ sentences for each translation from $\mathcal{TC}(w_1^S)$. In total, we have designed 360 sentences for each language pair (ES/IT/NL-EN). This is our *CWT+Wiki* evaluation dataset. Similarly as with CWT+JA-BNC, we have used 5 extra nouns with 20 sentences each as a development set to tune the parameters of our models. We use CWT+Wiki to evaluate our context-sensitive models of cross-lingual similarity in the task of the ES/IT/NL-to-EN word translation in context. Tab. 10.2 presents a small sample from both CWT+JA-BNC and CWT+Wiki evaluation datasets, and briefly illustrates the task of word translation in context.

**Evaluation Procedure.** Our task is to present the system a list of possible translations and let the system decide a *single most likely translation* given the word and its sentential context. Ground truth thus contains one word, that is, one correct translation for each sentence from the evaluation dataset. For CWT+JA-BNC, a word that was annotated by human annotators as the most

likely translation has been chosen as the correct translation [169]. In cases when two or more translations are equally likely, the system is rewarded if it proposes any of them as the correct translation. For CWT+Wiki, we have manually annotated the correct translation for the ground truth by inspecting the discourse in Wikipedia articles and the interlingual Wikipedia links. We measure the performance of all models again as *Top 1* accuracy ($Acc_1$). In this task, it denotes the number of word instances from the evaluation dataset whose top proposed candidate in the ranked list of translation candidates from $\mathcal{TC}$ is exactly the correct translation for that word instance as given by ground truth over the total number of test word instances (1,000 for CWT+JA-BNC and 360 for CWT+Wiki).

**Parameters.** We have tuned $\lambda_1$ and $\lambda_2$ on the development sets consisting of 5 nouns with 20 sentences each for both CWT+JA-BNC and CWT+Wiki. For all language pairs in CWT+JA-BNC we set $\lambda_1 = \lambda_2 = 0.7$, while we set $\lambda_1 = \lambda_2 = 0.9$ for all language pairs in CWT+Wiki. We use sorted context sets (see sect. 10.3.2) and perform a cut-off at $M = 3$ most descriptive context words in the sorted context sets for all models. In the following section we discuss the utility of this context pruning, as well as its influence on the overall results.

**Compared Models.** We test the performance of our DIRECT-FUSION, SMOOTHED-FUSION and LATE-FUSION models on both evaluation datasets. Since our main goal is to show the utility of context-aware word representations and context-sensitive models of cross-lingual similarity as opposed to out-of-context word representations and context-insensitive models, we compare their results with the context-insensitive models described in chapter 6 (NO-CONTEXT). For all compared models, we provide results with two different similarity functions discussed in chapter 6: (1) We have tested different SF-s again (e.g., the Kullback-Leibler and the Jensen-Shannon divergence, the cosine similarity) on the $K$-dimensional vector representations, and have detected that in general the best scores are obtained with the Bhattacharyya coefficient (BC), (2) We also adapt the Cue model (see eq. (10.6)) and report the results.

## 10.6 Experiments, Results and Discussion

In this section, we report the results accompanied with key conclusions in the task of word translation in context for all three language pairs. First, we display the results on the CWT+JA-BNC test dataset. Following that, we report the performance of our cross-lingual models of similarity in context on the CWT+Wiki dataset, and finally analyze the influence of context sorting and pruning.

### 10.6.1 Experiment I: Results on the CWT+JA-BNC Test Set

The performance of all models of cross-lingual semantic similarity on the CWT+JA-BNC test dataset is displayed in tab. 10.3 and tab. 10.4. Tab. 10.3 shows the results for Spanish-English and Italian-English, while tab. 10.4 shows the results for Dutch-English with two different training corpora (without and with Europarl training data). These results lead us to several conclusions:

(i) In general, the proposed models of contextualized semantic similarity which are consequently able to provide word translations in context outperform context-insensitive models of similarity that are able to produce only word translations in isolation (e.g., we observe an average increase of 9.1% for the BC+SMOOTHED FUSION combination, 8.7% for the BC+LATE-FUSION combination, 15.8% for Cue+SMOOTHED-FUSION, and 13.5% for Cue+LATE-FUSION). The improvements in results when taking context into account are observed for all three language pairs. Larger improvements are observed when using SMOOTHED-FUSION and LATE-FUSION which make an explicit distinction between the observed source word and its context in modeling (unlike the DIRECT-FUSION model which blends the observed source word and its context directly).

(ii) We have additionally compared our context-aware models with another context-insensitive model reported in [169]. This baseline model chooses the most probable translation from a word translation table obtained by an automated word alignment process on the sentence-aligned Europarl corpus. Lefever and Hoste [169] used GIZA++ [229] for word alignment and obtained the lists of most probable word translations. We make use of their lists in our experiments. The results of this baseline model are 0.435 for English-to-Spanish, 0.349 for English-to-Italian, and 0.332 for English-to-Dutch. It is very interesting to note that our context-aware models that rely on out-of-domain comparable corpora outperform even this context-insensitive model that was trained on the sentence-aligned parallel corpus that was utilized to induce sense inventories for the test dataset.

### 10.6.2 Experiment II: Results on the CWT+Wiki Test Set

The performance of all models of cross-lingual semantic similarity on the CWT+Wiki test dataset is displayed in tab. 10.5 and tab. 10.6. Again, tab. 10.5 displays the results for Spanish-English and Italian-English, while tab. 10.6 provides the results for Dutch-English with two different training corpora. Now, based on the results obtained on both test datasets presented in tables 10.3, 10.4, 10.5, 10.6, we observe several phenomena:

(i) We again observe improvements in the task of word translation in context when we employ context-sensitive models of cross-lingual similarity in place of

| **Direction:** | **EN→ES** | | **EN→IT** | |
|---|---|---|---|---|
| **Model** | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) |
| NO-CONTEXT | 0.434 | 0.449 | 0.331 | 0.335 |
| DIRECT-FUSION | 0.424 | 0.451 | 0.351 | 0.332 |
| SMOOTHED-FUSION | 0.472 | **0.504** | 0.357 | **0.386** |
| LATE-FUSION | **0.479** | 0.492 | **0.366** | 0.375 |

Table 10.3: Results on the CWT+JA-BNC test dataset. Training corpus is Wiki. Translation direction is EN-ES/IT.

| **Direction:** | **EN→NL** (Wiki) | | **EN→NL** (Wiki+EP) | |
|---|---|---|---|---|
| **Model** | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) |
| NO-CONTEXT | 0.298 | 0.289 | 0.315 | 0.322 |
| DIRECT-FUSION | 0.286 | 0.288 | 0.335 | 0.373 |
| SMOOTHED-FUSION | **0.331** | **0.334** | **0.341** | **0.387** |
| LATE-FUSION | 0.326 | 0.328 | 0.328 | 0.383 |

Table 10.4: Results on the CWT+JA-BNC test dataset displaying the difference in results when training on Wiki and Wiki+EP. Translation direction is EN-NL.

context-insensitive models. The improvements in scores are even more prominent on the CWT+Wiki dataset (e.g., we observe an average increase of 51.6% for the BC+DIRECT FUSION combination, 64.3% for BC+SMOOTHED-FUSION, 64.9% for BC+LATE-FUSION, 49.1% for Cue+DIRECT-FUSION, 76.7% for Cue+SMOOTHED-FUSION, and 64.5% for Cue+LATE-FUSION). Furthermore, since we observe improvements on both datasets for all three language pairs, we may conclude that the utility of the proposed probabilistic framework is not limited to only one translation direction, and it shows its potential for a variety of language pairs. In addition, the framework seems to be fairly robust as it displays a similar behavior on two different datasets.

(ii) The choice of a similarity function influences the results on both test datasets in both translation directions. On average, the Cue method as SF outperforms other standard similarity functions (e.g., Kullback-Leibler, Jensen-Shannon, cosine, the Bhattacharyya coefficient) in this evaluation task. However, it is important to state that *regardless of the actual choice of SF, context-aware models that modulate out-of-context word representations using the knowledge of local context outscore context-insensitive models that utilize non-modulated out-of-context representations* (with all other parameters equal).

| **Direction:** | **ES→EN** | | **IT→EN** | |
|---|---|---|---|---|
| **Model** | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) |
| NO-CONTEXT | 0.406 | 0.406 | 0.408 | 0.408 |
| DIRECT-FUSION | 0.617 | 0.575 | 0.714 | 0.697 |
| SMOOTHED-FUSION | 0.664 | **0.703** | 0.731 | **0.789** |
| LATE-FUSION | **0.675** | 0.667 | **0.742** | 0.728 |

Table 10.5: Results on the CWT+Wiki test dataset. Training corpus is Wiki. Translation direction is ES/IT-EN.

| **Direction:** | **NL→EN** (Wiki) | | **NL→EN** (Wiki+EP) | |
|---|---|---|---|---|
| **Model** | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) | $Acc_1$ (SF=BC) | $Acc_1$ (SF=Cue) |
| NO-CONTEXT | 0.433 | 0.433 | 0.433 | 0.433 |
| DIRECT-FUSION | 0.603 | 0.592 | 0.606 | 0.636 |
| SMOOTHED-FUSION | **0.669** | **0.712** | **0.692** | **0.761** |
| LATE-FUSION | 0.667 | 0.644 | 0.683 | 0.722 |

Table 10.6: Results on the CWT+Wiki test dataset displaying the difference in results when training on Wiki and Wiki+EP. Translation direction is NL-EN.

(iii) Results obtained on CWT+Wiki are typically higher than results obtained on CWT+JA-BNC, and the overall gain of context-aware models over context-insensitive models is higher on CWT+Wiki. There are several reasons: (1) The multilingual topic model was trained on comparable Wikipedia data. In case of CWT+Wiki, we again test on in-domain Wikipedia sentences. Note that the test sentences were not used as training data. In case of CWT+JA-BNC we test on sentences derived from another out-of-domain corpus. (2) Lists of translations for CWT+JA-BNC typically contain more translation candidates and are finer-grained. It is more difficult to capture slight translational variations with contextual variations in CWT+JA-BNC. For instance, it is more difficult to decide whether to translate the English word *post* as *correo*, *servicio postal*, *puesto*, *cargo*, *posición*, *correo*, *postal* in Spanish since all these translation candidates share a lot of contextual information. On the other hand, when translating Spanish word *gota* to English, it is much easier to decide whether a correct translation is *gout* or *drop*, since the contextual information clearly distinguishes the two possible meanings. (3) Due to its inherent encyclopaedic nature, Wikipedia sentences are typically more structured and informative than the sentences from CWT+JA-BNC (see again examples from tab. 10.2) which consequently leads to more reliable and more informative contextual knowledge

that is effectively exploited by our models.

(iv) The DIRECT-FUSION model, conceptually similar to a model of word similarity in context in monolingual settings [76], is again outperformed by the other two models, although the DIRECT-FUSION exhibits a much better performance on the CWT-Wiki test dataset. In DIRECT-FUSION, the observed word and its context are modeled in the same way, that is, the model does not distinguish between the word and its surrounding context when it computes the modulated probability scores $P'(z_k|w_1^S)$ (see eq. (10.9)). Unlike DIRECT-FUSION, the modeling assumptions of SMOOTHED-FUSION and LATE-FUSION provide a clear distinction between the observed word $w_1^S$ and its context $Con(w_1^S)$ and combine the out-of-context representation of $w_1^S$ and its contextual knowledge into a smoothed LM-inspired probabilistic model. As the results reveal, that strategy leads to better overall scores. The best scores in general are obtained by the SMOOTHED-FUSION model, but it is also outperformed by LATE-FUSION in several experimental runs where BC was used as SF. However, the difference in results between SMOOTHED-FUSION and LATE-FUSION in the experimental runs where LATE-FUSION outperforms SMOOTHED-FUSION is not statistically significant according to a chi-squared significance test ($p < 0.05$).

(v) The results for Dutch-English on both test datasets are influenced by the quality of training data. The performance of our models of similarity is higher for models that rely on latent-cross lingual topics estimated from the data of higher quality (i.e., compare the results when trained on Wiki and Wiki+EP in tab. 10.4 and tab. 10.6). The overall quality of our models of similarity is of course dependent on the quality of the latent cross-lingual topics estimated from training data, and the quality of these latent cross-lingual concepts is further dependent on the quality of multilingual training data (see also a similar finding for the task of bilingual lexicon extraction in chapter 8 and also in [314]).

(vi) Although Dutch is regarded as more similar to English than Italian or Spanish, we do not observe any major increase in the results on both test datasets for the English-Dutch language pair compared to English-Spanish/Italian. That phenomenon may be attributed to the difference in size and quality of our training Wikipedia datasets (e.g., see again the discussion in [314], or sect. 8.4 in chapter 8). Moreover, while the probabilistic framework proposed in this chapter is completely language pair agnostic as it does not make any language pair dependent modeling assumptions, we acknowledge the fact that all three language pairs comprise languages coming from the same phylum, that is, the Indo-European language family. Future extensions of our probabilistic modeling framework also include porting the framework to other more distant language pairs that do not share the same roots nor the same alphabet (e.g., English-Chinese/Hindi).

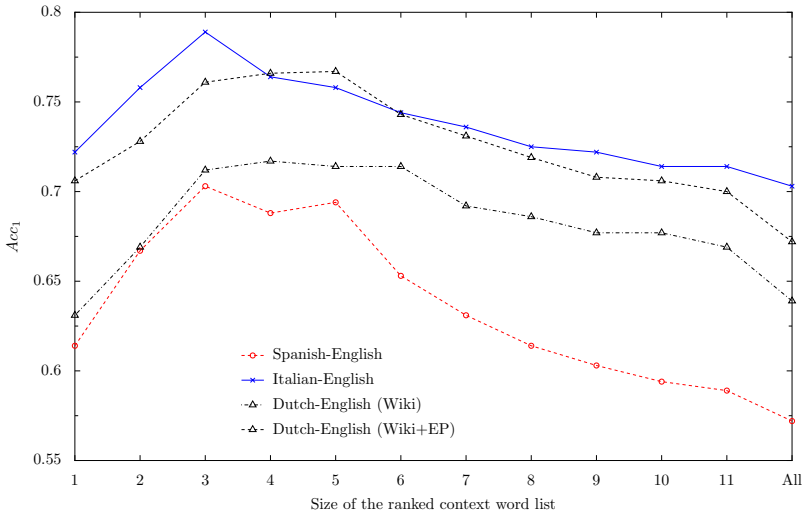



Figure 10.3: The influence of the size of sorted context on the accuracy of word translation in context. Test dataset is CWT+Wiki. The model is SMOOTHED-FUSION (SF=Cue).

### 10.6.3 Experiment III: Analysis of Context Sorting and Pruning

We also investigate the utility of context sorting and pruning and its influence on the overall results in the task of word translation in context. Therefore, we have conducted experiments with sorted context sets that were pruned at different positions, ranging from 1 (only the most similar word to $w_1^S$ in a sentence is included in the context set $Con(w_1^S)$) to *All* (all words occurring in a same sentence with $w_1^S$ are included in $Con(w_1^S)$). The monolingual similarity between $w_1^S$ and each potential context word in a sentence has been computed using BC on their out-of-context representations in the latent semantic space spanned by cross-lingual topics. Fig. 10.3 shows how the size of the sorted context influences the overall results on the CWT+Wiki test dataset. Tab. 10.7 reveals the increase in results when we utilize only a few best scoring words in a sentence as context words instead of the entire sentential context. The presented results have been obtained by the SMOOTHED-FUSION+Cue combination, but similar behavior is observed when employing other combinations.

The results clearly indicate the importance of context sorting and pruning. With that procedure we ensure that only the most semantically similar words in a given scope (e.g., a sentence) influence the choice of a correct word translation. In other words, closely semantically similar words in the same sentence are more reliable indicators for the most probable word meaning. They are more important

| | **ES→EN** | **IT→EN** | **NL→EN** (Wiki) | **NL→EN** (Wiki+EP) |
|---|---|---|---|---|
| **M** | $Acc_1$ | $Acc_1$ | $Acc_1$ | $Acc_1$ |
| 0 | 0.406 | 0.408 | 0.433 | 0.433 |
| 1 | 0.614 | 0.722 | 0.631 | 0.706 |
| 3 | **0.703** | **0.789** | 0.712 | 0.761 |
| 5 | 0.694 | 0.758 | **0.717** | **0.767** |
| 8 | 0.614 | 0.725 | 0.686 | 0.719 |
| 10 | 0.594 | 0.714 | 0.677 | 0.706 |
| All | 0.572 | 0.703 | 0.639 | 0.672 |

Table 10.7: Results for different sizes of sorted context sets. Test dataset is CTW+Wiki. The model is SMOOTHED-FUSION (SF=Cue).

and more informative in modulating the out-of-context word representations in context-sensitive models of similarity. For instance, given an occurrence of the word *coach* in a sentence (see tab. 10.2), the context word *team* bears more contextual information than the words *suspect* or *position*, and it is a much stronger indicator that the correct translation should be *allenatore* and not some other translation candidate.

By pruning the context we decrease the noise coming from words that co-occur with the given word in a sentence, but are not closely semantically related to the given word. These words do not bear any useful contextual information, and in some cases (e.g., if these words are very frequent in a training corpus from which the latent concepts are estimated) they might even negatively affect the modulation of out-of-context word representations by contextual knowledge.

## 10.7 Conclusions and Future Work

In this chapter, we have further extended our framework for modeling cross-lingual semantic similarity. We have described an extension of the framework which models cross-lingual semantic similarity in context and have presented its utility in the task of word translation in context. The key idea in this new context-sensitive approach is to represent words, regardless of their actual language, as distributions over latent cross-lingual topics/concepts, and both out-of-context and contextualized word representations are then presented in the same latent space spanned by the latent topics. A change in word meaning after observing its context is reflected in a change of its distribution over the latent topics.

Results on two evaluation datasets for three language pairs have shown the importance of the newly developed modulated or "contextualized" word representations in the task of word translation in context. As another contribution, we have further illustrated the benefit of using only the most informative contextual information in the contextualized models. The sorting and pruning of the context is based on the semantic word similarity in the same latent space spanned by the same latent cross-lingual topics/concepts. We have shown how to utilize the models of monolingual semantic similarity as an extra source of evidence in the cross-lingual models of similarity.

As a "by-product" of our work, we have constructed a new dataset that is suitable for testing the quality of cross-lingual models of semantic similarity in (sentential) context. Unlike previous datasets that always tackled the $English \rightarrow X$ direction in various related tasks, our dataset introduces the $X \rightarrow English$ direction and provides a small repository of highly ambiguous words in languages other than English (i.e., $X$ = Spanish, Dutch, Italian). The dataset may be easily extended with additional test instances and language pairs in future work.

In this chapter, we have introduced the core modeling premises, intuitions and a mathematical foundation behind the framework that relies on latent cross-lingual topics for modeling cross-lingual similarity in context. The proposed framework unfolds a series of new research questions and perspectives. Therefore, the paths of future work are manifold. Since the proposed framework is completely language pair agnostic and does not rely on any language pair specific knowledge in modeling, a straightforward path of future work is applying the framework to more language pairs (see the discussion in sect. 10.6.2). Additionally, one may further examine the influence of context scope (see sect. 10.3.2) on the contextualized models and study the behavior of the context-sensitive models of similarity when dealing with paragraphs or entire documents as contexts (see again the discussion in sect. 10.3.1). One may also study other methods of context ranking and pruning in order to capture only the most relevant context (see sect. 10.6.3). It is also worth studying the other models that induce latent semantic concepts from multilingual data (e.g., [117, 135, 333, 64, 91]) within this framework of context-sensitive models of cross-lingual similarity (see sect. 4.3 and sect. 4.6 from chapter 4). Moreover, we may apply the same modeling principle and "contextualize" our response-based second-order similarity model described in chapter 8. One may also investigate a similar approach to context-sensitive models of similarity that could operate with explicitly defined concept categories [94, 95, 49, 121, 122, 193]. Moreover, while in this work contextual features are co-occurring words and syntactic information is neglected, it is possible to explore contextualized models that rely on other contextual features, and other criteria of context aggregation and selection. For instance, similar to the model from Ó Séaghdha and Korhonen [228] in the monolingual setting, one

may introduce dependency-based contexts [231] instead of purely bag-of-words sentential and window-based contexts [17], and incorporate the syntax-based knowledge in the cross-lingual setting.

This chapter concludes our contributions to the field of distributional semantics. In part III we have proposed and described a new statistical framework for modeling cross-lingual semantic similarity which relies on the knowledge of latent cross-lingual topics (or more generally - latent cross-lingual concepts). The latent cross-lingual topics may be directly estimated from comparable data such as aligned Wikipedia articles given in different languages. Consequently, the statistical framework for modeling cross-lingual semantic similarity does not make any additional language-pair dependent assumptions, that is, it does not rely on an external bilingual lexicon, orthographic clues or predefined ontology/category knowledge, and it does not require parallel data. Moreover, *while the focus in this thesis is on cross-lingual models similarity that we deem more general, all the models described in this part are fully operational in simpler monolingual settings.* The nature of presentation in part III regards monolingual models of semantic similarity (both out of context and in context) that rely on latent cross-lingual topics only as special degenerate cases of the cross-lingual models which operate with only one language.

In part IV, the thesis moves its focus to another domain as we test the utility of latent cross-lingual topics in cross-lingual information retrieval. Similar to our statistical framework for modeling cross-lingual semantic similarity discussed in part III, in part IV we will propose and thoroughly describe a new statistical probabilistic framework for building cross-lingual information retrieval models. The framework is again supported by the knowledge of latent cross-lingual topics, and it again operates in the minimalist cross-lingual setting which requires only comparable data for training. In addition, in part IV we will demonstrate how to embed the knowledge of semantically similar words in our new cross-lingual retrieval models built within the framework.

## 10.8 Related Publications

[1] **I. Vulić** and M.-F. Moens. "Probabilistic models of cross-lingual semantic similarity based on latent cross-lingual concepts," submitted to the *19th Conference on the Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 25-29 October 2014, pp. xx-xx, ACL, 2014.

[2] **I. Vulić** and M.-F. Moens. "A probabilistic framework for modeling cross-lingual semantic similarity in context based on latent cross-lingual concepts," journal article in preparation for *Journal of Artificial Intelligence Research*, 2014.

# Part IV

# Cross-Lingual Information Retrieval

# Outline of Part IV

In part IV, the focus of the thesis shifts from NLP/CL to another related domain, as we present how to tackle the fundamental problem of information retrieval (IR) with an emphasis on cross-lingual information retrieval (CLIR). In CLIR the language of a user's query differs from the language of the target document collection. In this part, we address requirement R3 and research question RQ4 (see sect. 1.1 in chapter 1). In short, we explore whether it is possible to construct robust and cheap unsupervised statistical algorithms for monolingual and cross-lingual information retrieval again without any external translation resources for a variety of language pairs. We show how to use the shallow semantic knowledge coming from the output distributions of a multilingual probabilistic topic model (see sect. 4.4.3 in chapter 4) in (CL)IR models. As a major contribution, part IV introduces, describes and thoroughly evaluates a new statistical probabilistic framework for (CL)IR which relies on the knowledge of latent cross-lingual topics/concepts which can be again estimated from non-parallel data without any additional linguistic knowledge. Part IV is logically divided into two chapters:

I. Chapter 11 provides a short introduction to IR and CLIR. Following that, it proposes and describes the framework for building MuPTM-based (CL)IR models, covers the key modeling assumptions, and introduces several new CLIR models, starting from a basic MuPTM-based CLIR model, and subsequently introducing more elaborate models.

II. Chapter 12 provides a short introduction to relevance modeling in IR and CLIR and then describes an extension of the proposed probabilistic framework and shows how to build more robust and more effective language pair agnostic retrieval models by combining the advantages of multilingual probabilistic topic modeling and relevance modeling.

# 11

# Multilingual Topic Models in (Cross-Lingual) Information Retrieval

*It is a very sad thing that nowadays there is so little useless information.*

— Oscar Wilde

## 11.1 Introduction

*Information retrieval (IR)* is the field of research that deals with finding documents in a large document collection[1] that are relevant to a user's needs. The user typically provides her/his *information need* in a form of a query that contains keywords or a short description of the information need, and then expects a retrieval of information that is relevant to the provided query. Formally, the basic IR setup follows these lines: the user's *query* $Q$ containing $m$ query terms, $Q = \{q_1, \ldots, q_m\}$ is issued to retrieve documents from a *target document collection* $\mathcal{DC}$ that contains $\mathcal{D} = |\mathcal{DC}|$ documents, $\mathcal{DC} = \{d_1, \ldots, d_{\mathcal{D}}\}$. Documents from the document collection are ranked according to their relevance to the query $Q$. Different models of document ranking have been proposed in the literature, and investigating the variety of models for relevance detection and ranking constitutes the core of theoretical IR research.

[1] More generally, we could talk about a large item collection where the items may be text documents, videos, pictures, audio sequences, etc. However, in this thesis we focus on text documents and text-based information retrieval.

In this chapter, we present how to effectively apply multilingual probabilistic topic models in building monolingual, cross-lingual and even multilingual information retrieval models. *Cross-lingual information retrieval (CLIR)* is the sub-field of IR where the query $Q$ is issued in a language different from the language of the target document collection $\mathcal{DC}$. As an extension of the CLIR setting to the setting where more than 2 languages are involved, *multilingual information retrieval (MIR)* is the sub-field of IR where the documents $d_j \in \mathcal{DC}$ may be provided in multiple different languages. A MIR model then has to rank relevant documents from the target collection, irrespective of their actual language. The use of multilingual probabilistic topic models in CLIR and MIR models is a completely new field of research that has been pioneered within the scope of this thesis. Furthermore, we have developed new state-of-the-art IR models for ad-hoc monolingual retrieval that rely on the shallow latent semantic knowledge obtained by probabilistic topic models. The text in this chapter mostly tackles the CLIR setting, but the reader has to be aware that all proposed models are fully functional for multilingual and monolingual retrieval (the relation between monolingual, cross-lingual and multilingual retrieval models is analogous to the relation between LDA, BiLDA and PolyLDA; see sect. 4.4).

In the standard CLIR setting, at the time of retrieval the query in the source language $Q^S$ is typically translated into the target language of the target document collection $\mathcal{DC}^T$ with the help of a machine-readable dictionary or a machine translation system [133, 104, 172]. Another translation direction is also possible, that is, translating all documents from the target collection to the language of the query. However, this method is less common due to its increased computational complexity. After both the query and the target collection are transferred into the same language, a myriad of monolingual retrieval techniques may be applied (e.g., [242, 22, 163]). Once a user has retrieved relevant documents for a particular query, they can be translated to the language of the user, possibly by means of manual translation in case resources for automatic translation are unavailable.

In this chapter, we address the question whether effective cross-lingual information retrieval models can be built in case machine-readable translation dictionaries or MT systems that are hand-built or extracted from large parallel sentence-aligned corpora are absent. A number of words might appear with the same meaning in different languages (especially when dealing with languages from the same family). However, when only using a monolingual retrieval model for CLIR, we will miss many relevant documents. Moreover, a word might exhibit the same orthography in different languages, but actually mean something different. Consequently, we need some kind of translation resource, preferably built automatically from less expensive and abundant non-parallel corpora. In case when readily available translation resources are unavailable,

based on their modeling properties, multilingual probabilistic topic models should serve as a valid tool to build a CLIR system that does not rely on any external translation resource and can be trained on general-domain non-parallel data (e.g., Wikipedia) and later inferred on and used on in-domain data (e.g., newswire corpora).

We investigate whether the transfer of a source language query into the target language may be accomplished by means of a multilingual probabilistic latent topic model that is embedded in the language modeling (LM) IR framework [53]. The language models for retrieval have a sound statistical foundation, can leverage statistical estimation to optimize retrieval parameters, and allow for a straightforward integration of complementary retrieval clues into a retrieval model. They can be easily adapted to complex retrieval tasks and have already shown their value in cross-lingual retrieval settings, e.g., by embedding translation probabilities obtained from a translation dictionary into the retrieval model. Our aim is to exploit the probability distributions over latent-cross lingual topics as a translation resource, since they provide a language-independent content representation of the documents.

The contributions of the work reported in this chapter are as follows. First, we propose a new statistical probabilistic framework for constructing CLIR models and introduce a set of new MuPTM-based CLIR models. Second, we demonstrate the applicability and usefulness of the MuPTM-induced lexicons (from part III) in the LM CLIR framework. Third, we successfully integrate the knowledge from the lexicons and the knowledge from probability distributions of a multilingual probabilistic topic model into a novel evidence-rich cross-lingual statistical retrieval model which uses only internal evidence, and perform a full-fledged evaluation and comparison of all our retrieval models for: (1) the simpler task of English-Dutch and Dutch-English known-item search performed on Wikipedia articles, and (2) the task of cross-lingual English-Dutch and Dutch-English ad-hoc information retrieval on the standard benchmarking CLEF test collections. We show that the results obtained by our retrieval models, which do not exploit any linguistic knowledge from an external translation resource are competitive with and sometimes even display a better performance than dictionary-based models for CLIR. Finally, we question the true meaning and validity of perplexity, a standard theoretical quantitative measure that is most commonly used to compare various latent topic models in vitro (see sect. 4.5). We demonstrate that better perplexity scores do not necessarily lead to better results in a "real-life" application of topic models such as cross-lingual information retrieval.

The chapter is structured as follows. We review related work in sect. 11.2. Following that, we describe our new probabilistic LM (CL)IR framework in sect. 11.3. Our experimental setup with evaluation tasks, training data, test

collections, queries, and evaluation metrics is discussed in sect. 11.4. We provide a thorough analysis of our retrieval framework in sect. 11.5, while sect. 11.6 lists main conclusions of the chapter.

## 11.2 Related Work

Cross-lingual information retrieval is a broad and well-studied research topic (e.g., [111, 224, 266, 223]). As mentioned, existing methods typically rely on a translation dictionary to bridge documents of different languages. In another typical setting, cross-lingual information are learned based on parallel corpora and correlations found in the paired documents [188], or are based on Latent Semantic Analysis (LSA) applied on a parallel corpus. In the latter case, a singular value decomposition is applied on the term-by-document matrix, where a document is composed of the concatenated text in two languages, and after rank reduction, the document and the query are projected into a lower-dimensional space [81, 179, 47, 328]. The term-by-document matrix formed by concatenated parallel documents was also used to generate probabilistic term translations using a standard monolingual PLSA and LDA. The probabilities are then used in a CLIR model [216, 261]. Our work follows this line of thinking, but uses multilingual probabilistic topic models trained on a comparable document-aligned corpus, which might be different from the document collection used for retrieval. In addition, our models are trained on the individual documents in different languages, but paired through the latent cross-lingual topical space and, due to that fact, we expect our models to lead to better results than CLIR models relying on standard algebraic models such as LSA or monolingual topic models such as PLSA or LDA. An LDA-based LM IR framework for monolingual ad-hoc retrieval is described in [323]. However, that framework is subsumed by our more general framework presented in this chapter.

Transfer learning techniques, where knowledge is transferred from one source to another, are also used in the frame of cross-lingual text classification and clustering. Transfer learning bridged by probabilistic topics obtained via PLSA was proposed in [328] for the task of cross-domain text categorization. Recently, knowledge transfer for cross-domain learning to rank the answer list of a retrieval task was described in [45], while Takasu [282] proposes cross-lingual keyword recommendation using latent topics.

The work conducted in this chapter of the thesis is the first application of the lexicons extracted by a multilingual topic model in a *real-life* task such as CLIR. The usage of semantically similar words from the ranked lists obtained by various models of cross-lingual semantic similarity (see part III) may be observed

as a query expansion technique, constructed to further improve the effectiveness of a CLIR model. Query expansion techniques relying on a statistical similarity measure among terms stored in an automatically generated thesaurus/lexicon are described in [268, 1], but the prior work differs from ours in both construction of the lexicon and its usage in the CLIR model (e.g., they employ non-probabilistic retrieval models).

## 11.3 MuPTM-Based CLIR

This section provides a theoretical insight into probabilistic LM cross-lingual information retrieval (CLIR) models that rely on per-topic word distributions and per-document topic distributions from sect. 4.4. We start from a basic model that relies only on the language-independent topical representations of documents, and then gradually build more elaborate CLIR models by embedding additional clues and representations in the probabilistic framework, including the knowledge from the MuPTM-induced bilingual lexicons from part III.

### 11.3.1 MuPTM-Basic CLIR Model

Given a target document collection $\mathcal{DC}^T$, that is, the set of $\mathcal{D} = |\mathcal{DC}^T|$ documents, $\mathcal{DC}^T = \{d_1^T, d_2^T, \ldots, d_D^T\}$ in a target language $L_T$, and a query $Q^S$ in a source language $L_S$, the task is to rank the documents according to their relevance to the query. The ranked list of documents in the target document collection given the query $Q^S$ is denoted as $DRank(Q^S)$. We follow the approach that relies on probabilistic language models (see sect. 2.3 in chapter 2) in monolingual information retrieval. We again utilize the *bag-of-words* assumption. It states that the ordering of words in a document is not important, and that the *conditional independence* assumption holds, that is, the words are conditionally independent given the documents. In this chapter, we opt for the *query likelihood* modeling paradigm where the score of each document $d_j^T$ is the likelihood of its model generating the query $Q^S$. The probability $P(Q^S|d_j^T)$ that the query $Q^S$ is generated from the document model $d_j^T$ is calculated based on the unigram language model which assumes the independence between the query terms. More formally, $P(Q^S|d_j^T)$ is calculated as follows:

$$P(Q^S|d_j^T) = P(q_1^S, \ldots, q_m^S|d_j^T) = \prod_{i=1}^{m} P(q_i^S|d_j^T) \tag{11.1}$$

The main difference between monolingual IR and CLIR and the main obstacle in the CLIR setting lies in the fact that documents are not given in the same language as the query. Therefore, one needs to find a way to effectively bridge the gap between the two involved languages, or the so-called "lexical chasm" [21]. The typical approach is to apply machine-readable translation dictionaries, translate the query and perform monolingual retrieval on the translated query. However, if a translation resource is absent or unavailable, one needs to find another solution. In lack of any other translation resource, we propose to exploit the shared latent cross-lingual topical space obtained by a multilingual probabilistic topic model, that is, to use the sets of the output per-topic word distributions and per-document topic distributions discussed in sect. 4.4. Combining eq. (4.6) and eq. (4.7), we can now rewrite eq. (11.1) by calculating the probability $P(q_i^S|d_j^T)$ in terms of the two MuPTM-related probability distributions:

$$P(q_i^S|d_j^T) = (1-\delta_1)\sum_{k=1}^{K} \overbrace{P(q_i^S|z_k)}^{\text{in source}} \underbrace{P(z_k|d_j^T)}_{\text{in target}} + \delta_1 P(q_i^S|Ref^S)$$

$$= (1-\delta_1)\sum_{k=1}^{K} \phi_{k,i}^S \theta_{j,k}^T + \delta_1 P(q_i^S|Ref^S) \tag{11.2}$$

$\delta_1$ is the interpolation parameter, while $P(q_i^S|Ref^S)$ is the maximum likelihood estimate of the query word $q_i^S$ in a monolingual source language reference collection $Ref^S$. It assigns a non-zero probability to words unobserved during the training of the topic model in case it occurs in the query. Here, we use the observation that latent cross-lingual topics span a latent language-independent space shared between the languages. If that observation holds, each document, regardless of its actual language, may be represented as a mixture of latent cross-lingual topics in that space. Furthermore, it is justified to use the per-topic word distributions for the source language to predict the probability that the word $q_i^S$ from the query $Q^S$ will be sampled from the latent cross-lingual topic $z_k$. By modeling the procedure of "sampling" of query terms given in language $L_S$ from the documents given in language $L_T$ through the shared latent cross-lingual space, we have established a link between $L_S$ and $L_T$ by means of MuPTM.

As mentioned before, we may run/infer the multilingual topic model on any monolingual collection in the source or the target language. Inferring the model in this context actually means learning the representation of each document in a collection as a mixture of latent cross-lingual topics as provided by the per-document topic distributions (see sect. 4.2.1, sect. 4.4.3 and sect. 4.4.4). Note that when operating with only one language, the model may be exploited for

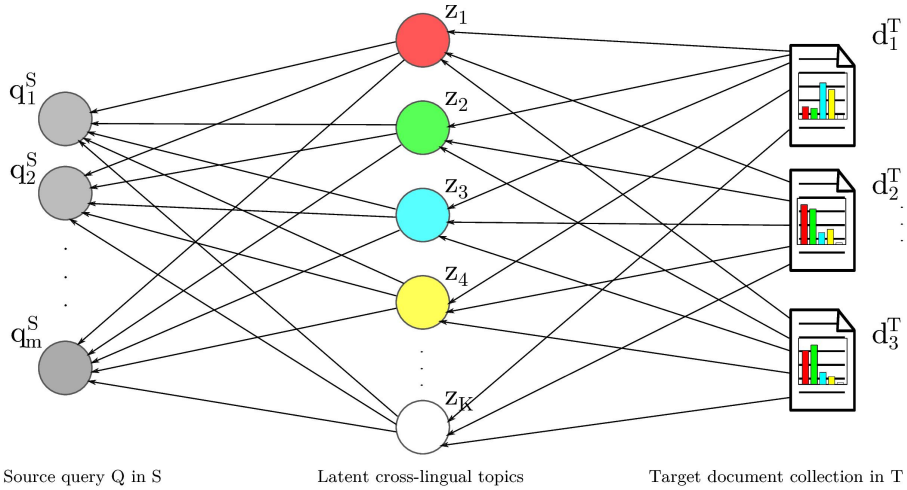



Figure 11.1: MuPTM-Basic retrieval model: An illustrative graphical presentation of the basic retrieval model that relies only on the latent layer of cross-lingual topics obtained by a multilingual probabilistic topic model.

monolingual retrieval and it is conceptually similar to the model from Wei and Croft [323]. Moreover, since documents have the same language-independent representation given by the distributions over cross-lingual topics, it allows for retrieving documents from a target collection given in multiple languages. In other words, documents relevant to the query may be in different languages, and the proposed model is able to process it in an uniform way. All further models will have the property of operating in the monolingual, cross-lingual and multilingual settings without any conceptual difference in the modeling approach. We can now merge all the steps into one coherent process to calculate the probability $P(Q^S = q_1^S, q_2^S, \ldots, q_m^S | d_j^T)$ for each $d_j^T \in \mathcal{DC}^T$, where $Q^S$ denotes a query in the source language, and $d_j^T$ denotes a document in the target document collection $\mathcal{DC}^T$. Alg. 11.1 presents an overview of the retrieval process.

This procedure outputs a ranked list $DRank(Q^S)$ of all documents in the target document collection according to their respective query likelihood scores. The intuitive graphical representation of this basic MuPTM-based CLIR technique that connects documents given in target language $L_T$ with query words given in source language $L_S$ is displayed in fig. 11.1. There, each target document is represented as a mixture of latent language-independent cross-lingual topics (colored bars which denote per-document topic distributions) and assigns a probability value $P(z_k|d_j^T)$ for each $z_k \in \mathcal{Z}$ (edges between documents and

topics). Moreover, each cross-lingual topic may generate each query word by the probability $P(q_i^S|z_k)$ given by per-topic word distributions (edges between topics and query words). Since this model constitutes the basic building block for all further more elaborate models that rely on the knowledge from the trained multilingual probabilistic topic models, we name this model *MuPTM-Basic*.

**Algorithm 11.1**: MuPTM-Basic retrieval model

**Input** : bilingual training corpus $\mathcal{C} = \mathcal{C}_S \cup \mathcal{C}_T$, target document collection $\mathcal{DC}^T$, user query $Q^S$ ;

1: **train** the model on a (usually general-domain) training corpus and learn per-topic word distributions $\phi$ and $\psi$, and per-document topic distributions ;

2: **infer** the trained model on $\mathcal{DC}^T$ in the target language $L_T$ and obtain per-document topic distributions $\theta^T$ for all documents in $\mathcal{DC}^T$ ;

3: **compute** the query relevance in the target collection:

**foreach** target document $d_j^T \in \mathcal{DC}^T$ **do**

  **foreach** query term $q_i^S \in Q^S$ **do**

    (a): **obtain** probabilities $\phi_{k,i} = P(q_i^S|z_k)$ from per-topic word distributions for $S$, for all $k = 1, \ldots, K$ ;

    (b): **obtain** probabilities $\theta_{j,k}^T = P(z_k|d_j^T)$, for all $k = 1, \ldots, K$ ;

    (c): **combine** the probabilities to obtain the final probability that a source term $q_i^S$ is generated by a document model $d_j^T$ via latent cross-lingual topics: $P_{muptm}(q_i^S|d_j^T) = \sum_{k=1}^{K} P(q_i^S|z_k)P(z_k|d_j^T)$ ;

  (d): **compute** the final query likelihood for the entire query:

$$P(Q^S|d_j^T) = \prod_{i=1}^{m} P(q_i^S|d_j^T) = \prod_{i=1}^{m} \left( (1-\delta_1) \sum_{k=1}^{K} \phi_{k,i}\, \theta_{j,k}^T + \delta_1 P(q_i^S|Ref^S) \right)$$

4: **rank** all documents $d_j^T \in \mathcal{DC}^T$ according to their respective scores $P(Q^S|d_j^T)$: $DRank(Q^S)$ ;

**Output**: $DRank(Q^S) \rightarrow$ the ranking of all documents from $\mathcal{DC}^T$ according to their relevance to $Q^S$ ;

## 11.3.2 MuPTM-DM CLIR Model

The MuPTM-Basic model from the previous section may be effectively combined with other models that capture additional evidence for estimating the probability $P(q_i^S|d_j^T)$. When dealing with monolingual retrieval, Wei and Croft [323] have detected that their model that relies on knowledge from a monolingual probabilistic topic model (e.g., LDA) is too coarse to be used as the only representation for retrieval and, consequently, to produce quality retrieval

results. Therefore, in the monolingual setting they have linearly combined it with the original document model (DM) that relies on a unigram language model and observed a major improvement in their results. We can follow the same principle in the cross-lingual setting, since a certain amount of content-bearing words from the query such as named entities does not change across languages. For instance, if the user is searching for the document about the volcano "Mauna Loa" in Croatian or Dutch, there is a fair chance that relevant documents in English, German, or even Finnish, Hungarian and the Basque language may be retrieved, since the query term "Mauna Loa" does not change over any of these languages. Therefore, we can combine the representation by means of a multilingual probabilistic topic model with the knowledge of the shared words across languages within the unified language modeling framework.

First, we provide a description of the model that relies only on the words shared across languages. Following that, we show how to combine the two different representations in a combined CLIR model.

**DM-Basic CLIR Model.** The probability $P(q_i^S|d_j^T)$ from eq. (11.1) may be estimated from a standard smoothed document model that relies on a relative frequency of a word in a document. We adopt the standard Dirichlet smoothing according to evaluations and findings from [332]. The Dirichlet smoothing acts as a length normalization parameter and penalizes long documents. The probability $P(q_i^S|d_j^T)$ is then:

$$P(q_i^S|d_j^T) = (1-\delta_2)\Big(\frac{N_{d_j^T}}{N_{d_j^T}+\mu}P_{mle}(q_i^S|d_j^T) + (1-\frac{N_{d_j^T}}{N_{d_j^T}+\mu})P_{mle}(q_i^S|\mathcal{DC}^T)\Big) + \delta_2 P(q_i^S|Ref^S) \tag{11.3}$$

$N_{d_j^T}$ denotes the length of the document $d_j^T$ computed as the number of word tokens in the document, $\mu$ is the parameter of the Dirichlet prior (see, e.g., [332]), $\delta_2$ is another interpolation parameter, and $P(q_i^S|Ref^S)$ is again the background probability of $q_i^S$, calculated over a large reference corpus $Ref^S$ given in the source language. It again provides smoothing by assigning a non-zero probability to words that have zero occurrences in the target collection. $P_{mle}(q_i^S|d_j^T)$ denotes the maximum likelihood estimate of the word $q_i^S$ in the document $d_j^T$, and it is computed as the relative term frequency of $q_i^S$ in $d_j^T$. On the other hand, $P_{mle}(q_i^S|\mathcal{DC}^T)$ denotes the maximum likelihood estimate of $q_i^S$ in the whole target collection $\mathcal{DC}^T$, which is again calculated as the relative *corpus frequency* of term $q_i^S$ in the corpus $\mathcal{DC}^T$. This probability acts as a smoothing parameter in the *collection smoothing* scheme where, according to the scheme, if a term is not available in a document, its probability should be close to the probability it has in the whole collection. These two probability

scores are calculated as follows:

$$P_{mle}(q_i^S|d_j^T) = \frac{tf_{d_j^T}(q_i^S)}{N_{d_j^T}} \qquad P_{mle}(q_i^S|\mathcal{DC}^T) = \frac{cf_{\mathcal{DC}^T}(q_i^S)}{\sum_{d_j^T \in \mathcal{DC}^T} N_{d_j^T}} \tag{11.4}$$

where $tf_{d_j^T}(q_i^S)$ counts the number of occurrences of $q_i^S$ in $d_j^T$, and $cf_{\mathcal{DC}^T}(q_i^S)$ counts the number of occurrences of $q_i^S$ in the entire document collection $\mathcal{DC}^T$.

Since this model is a simple retrieval model that estimates the query likelihood score using a document model that relies only on relative word frequencies as clues for retrieval, and does not rely on any other representation, we name it the *DM-Basic* model.

**Combining the Models.** We are now able to combine the *MuPTM-Basic* model from sect. 11.3.1 with the *DM-Basic* model using a simple linear interpolation, that is, the Jelinek-Mercer smoothing, again following [332]:

$$P(q_i^S|d_j^T) = \lambda P_{dm}(q_i|D_J) + (1-\lambda)P_{muptm}(q_i|D_J) \tag{11.5}$$

$$\begin{aligned} P(q_i^S|d_j^T) = \lambda\Big((1-\delta_2)\big(\frac{N_{d_j^T}}{N_{d_j^T}+\mu}P_{mle}(q_i^S|d_j^T) \\ + (1-\frac{N_{d_j^T}}{N_{d_j^T}+\mu})P_{mle}(q_i^S|\mathcal{DC}^T)\big) + \delta_2 P(q_i^S|Ref^S)\Big) \\ + (1-\lambda)P_{muptm}(q_i^S|d_j^T) \end{aligned} \tag{11.6}$$

where $P_{muptm}$ denotes the MuPTM-Basic model described by eq. (11.2), $P_{dm}$ denotes the DM-Basic model given by eq. (11.3), and $\lambda$ is the interpolation parameter. We call this final combined model the *MuPTM-DM* model. The combined model presented here is straightforward, since it directly uses words shared across two languages. One might also use cognates, that is, orthographically similar words identified, for instance, with the *edit distance* metric (see, e.g., [218]) instead of the shared words only.

### 11.3.3 MuPTM-SemLex CLIR Model

Now, we may continue embedding additional clues into the probabilistic language modeling framework for retrieval. As the next step, we demonstrate how to exploit the knowledge from the models of monolingual and cross-lingual semantic similarity when building new probabilistic (CL)IR models.

**Another Clue For Retrieval: Semantically Similar Words.** Recall that all the models of cross-lingual semantic similarity that have been discussed in part III generate ranked lists of semantically similar words, where synonymy is not the only observed relation of semantic similarity. In chapter 6 (see sect.

| (1) **vlucht** (flight) | (2) **reclame** (advertisement) | (3) **munt** (currency) |
|---|---|---|
| airlines | advertising | currency |
| airline | advertisements | currencies |
| carriers | placement | parities |
| overbooked | advertisers | fluctuation |
| easyjet | advertisement | devaluations |
| frills | stereotyping | euro |
| flights | billboards | devaluation |
| booking | adverts | overvalued |
| booked | advert | peseta |
| ryanair | advertise | fluctuations |

Table 11.1: Lists of the top 10 translation candidates (Dutch to English), where the correct translation is not found (column 1), lies hidden lower in the list (2), and is retrieved as the first candidate (3). Obtained with the TI+Cue method of cross-lingual semantic similarity.

6.2.1), we have defined that for some word $w_1^S$, $RL(w_1^S)$ consists of all $w_j^T \in V^T$ ranked according to their respective similarity scores $sim(w_1^S, w_j^T)$. Additionally with $RL_M(w_1^S)$ we have denoted the ranked list that is pruned at position $M$ and thus contains only the top $M$ words that are semantically similar to $w_1^S$. Such lists provide comprehensible and useful contextual information for the source word, even when the correct translation candidate is absent in the top $M$ candidates, as presented in tab. 11.1. We will show that the models of cross-lingual semantic similarity serve as a useful aid for CLIR models. We can easily turn the pruned ranked list $RL_M(w_1^S)$ into a probabilistic semantic lexicon that can then be integrated in the retrieval process. The probability $P(w_j^T|w_1^S)$, which models the degree of semantic similarity between a word $w_1^S$ in the source language and a word $w_j^T$ in the target language that occurs in the pruned list $RL_M(w_1^S)$ is calculated as follows:

$$P(w_j^T|w_1^S) = \frac{sim(w_1^S, w_j^T)}{\sum_{l=1}^{M} sim(w_1^S, w_l^T)} \tag{11.7}$$

The probability scores $P(w_j^S|w_1^T)$ in the opposite translation direction may be computed in the exact same fashion. The probability scores $P(w_j^T|w_1^S)$ and $P(w_j^S|w_1^T)$ are dependent on the value of $M$. We use the pruned ranked lists to

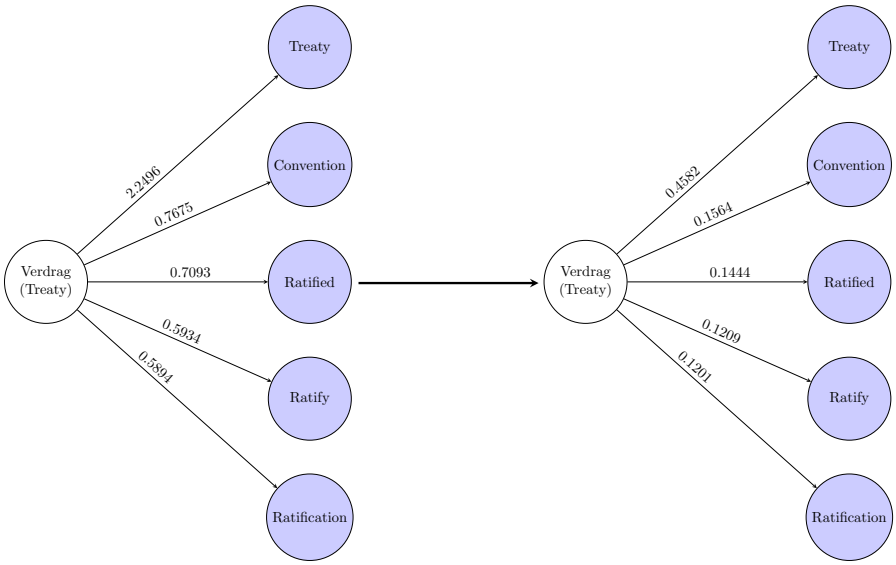



Figure 11.2: An example of a probabilistic semantic lexicon entry from a Dutch-English lexicon obtained from top $M = 5$ words from the ranked list. The scores on the edges on the left side are unnormalized similarity scores (the higher the score, the higher their semantic similarity). The scores on the edges on the right side (after the thick arrow) present normalized probability scores $P(w_j^T|w_1^S)$ after eq. (11.7) is employed.

decrease the computational complexity and to remove noise coming from words that do not exhibit any semantic relation to $w_1^S$.

**SemLex-Basic CLIR Model.** The simple model that uses the knowledge from the probabilistic lexicon relies on eq. (11.3). In case a source word $q_i^S$ is shared across languages and exists in the target language vocabulary $V^T$, eq. (11.3) is applied directly. On the other hand, if the source word $q_i^S$ does not exist in the target vocabulary, we need to reach out for the probabilistic semantic lexicon. We closely follow the translation model as presented in [22, 327]. If top $M$ semantically similar words from the probabilistic lexicon entry are taken into account for retrieval, the probability $P(q_i^S|d_j^T)$ is then:

$$P(q_i^S|d_j^T) = (1-\delta_3)\sum_{l=1}^{M} P(q_i^S|w_l^T)P(w_l^T|d_j^T) + \delta_3 P(q_i^S|Ref^S) \tag{11.8}$$

The summation goes over the top $M$ target words from the ranked list $RL_M(q_i^S)$. $P(q_i^S|w_l^T)$ is a translation probability for the words $q_i^S$ and $w_l^T$ from the entry calculated by eq. (11.7) when only top $M$ words are taken into account, while

$P(w_l^T|d_j^T)$ is computed as the first term of eq. (11.3) (preceded by $(1-\delta_2)$ in that equation). $P(q_i^S|Ref^S)$ is the background probability that is needed in case when there is no lexicon entry for the query word $q_i^S$. Since the model utilizes only the evidence from the probabilistic semantic lexicon combined with the evidence of shared words, and therefore constitutes a basic CLIR model that relies on the knowledge of semantically similar words, we name this model the *SemLex-Basic* model. Note that the model specified by eq. (11.8) allows for integrating any probabilistic lexicon, so even the external translational resources such as machine-readable dictionaries or lexicons acquired from parallel data are trivially integrated into the LM (CL)IR framework.

**The Final Model: MuPTM-SemLex.** The next model combines the knowledge from the probabilistic semantic lexicon as exploited in the SemLex-Basic model from the previous section with the knowledge coming from latent cross-lingual topics as exploited in the MuPTM-Basic model in sect. 11.3.1. This new model closely follows eq. (11.6), but instead of the DM-Basic model that was previously utilized to model the probability $P_{uni}(q_i^S|d_j^T)$ (see eq. (11.5)), it now utilizes the procedure of the SemLex-Basic model to estimate the probability score $P_{uni}(q_i^S|d_j^T)$ in eq. (11.5) and eq. (11.6). Since this model is realized as a combination of MuPTM-Basic and SemLex-Basic, we name this model the *MuPTM-SemLex* model.

The model is supported by several *a priori* assumptions:[2] (1) If a word occurs in both source and target language vocabularies, it is reasonable to assume that the word speaks for itself more than its translations do (for instance, if someone is searching for documents related to *Angela Merkel* or *Barack Obama*, no translation is needed), (2) If a word is not shared, one may use a list of semantically similar words in the target language, from the probabilistic semantic lexicon obtained from a training corpus and based on per-topic word distributions learned during the training of a multilingual topic model. Using a fully corpus-based probabilistic semantic lexicon is convenient, since it uses the same infrastructure as the multilingual topic model used to obtain topical representations of documents in $\mathcal{DC}^T$. Hence the model does not require any additional translational resource nor dictionary, (3) The "MuPTM-part" of the model, where by the "MuPTM-part" of the CLIR model we assume the part of the model given by eq. (11.2), introduces additional topical knowledge, since it connects words in the source language with documents in the target language through the shared latent space of cross-lingual topics and clusters/groups words

[2]It is straightforward to construct another variant of the MuPTM-SemLex model that provides lists of semantically similar words even for words that are shared across languages. In that case, a query expansion by means of semantically similar words is carried out for every query term, but since the two models exhibit comparable performances, we present only the results obtained with the first variant of the MuPTM-SemLex model.

appearing in similar contexts. The retrieval process with the MuPTM-SemLex model is summarized in alg. 11.2.

## 11.4 Experimental Setup

### 11.4.1 Evaluation Tasks: Known-Item Search and Ad-Hoc Search

The first evaluation task is a simulation of the *known-item search*. The known-item search is an important information finding activity which was long central to research and application in library and information sciences [166], but it has recently also gained a lot of attention in the information retrieval community [230]. In short, in this task the user knows that a particular target document exists and remembers its partial content, but does not know where to find the actual document. Example "document" types may be a Web page, a Wikipedia article, a report, an ordinary text document, or even an audio or a video file or a sequence. The known-item search is opposed to *subject or ad-hoc search* in which unknown documents, which may satisfy an information need are being searched for.

### 11.4.2 Training Collections

We work with the English-Dutch (EN-NL) language pair, and utilize two sub-corpora. The data used for training of the multilingual topic model is the same as before (see, e.g., sect. 8.3 in chapter 8). The first corpus consists of 7, 612 paired Wikipedia articles in English and Dutch (Wiki), while the second sub-corpus consists of the 6, 206 aligned English-Dutch Europarl documents [153] (EP) (see sect. 3.2.1 and sect. 3.2.2 in chapter 3. Again, although Europarl is a parallel corpus, no explicit use is made of sentence-level alignments, and we treat it only as a document-aligned corpus.

Instead of operating only with nouns as before (see sect. 8.3), as a preprocessing step we have removed only stop words (429 in English and 110 in Dutch), and words that occur less than 5 times in the complete training corpus. Our final vocabularies consist of 76, 555 words in English, and 71, 168 Dutch words. The final combined corpus is labeled as Wiki+EP.

**Algorithm 11.2**: MUPTM-SEMLEX RETRIEVAL MODEL

**Input** : bilingual training corpus $\mathcal{C} = \mathcal{C}_S \cup \mathcal{C}_T$, target document collection $\mathcal{DC}^T$, user query $Q^S$ ;

1: **train** the model on a (usually general-domain) training corpus and learn per-topic word distributions $\phi$ and $\psi$, and per-document topic distributions ;

2: **infer** the trained model on $\mathcal{DC}^T$ in the target language $L_T$ and obtain per-document topic distributions $\theta^T$ for all documents in $\mathcal{DC}^T$ ;

3: **compute** the query relevance in the target collection:

**foreach** target document $d_j^T \in \mathcal{DC}^T$ **do**

  **foreach** query term $q_i^S \in Q^S$ **do**

    I: **calculate** the probability $P_{dm}(q_i^S|d_j^T)$:

    **if** the source word $q_i^S$ is present in the target vocabulary $V^T$ **then**

$$P_{dm}(q_i^S|d_j^T) = (1-\delta_2)\Big(\frac{N_{d_j^T}}{N_{d_j^T}+\mu}P_{mle}(q_i^S|d_j^T) + \Big(1-\frac{N_{d_j^T}}{N_{d_j^T}+\mu}\Big)P_{mle}(q_i^S|\mathcal{DC}^T)\Big) + \delta_2 P(q_i^S|Ref^S)$$

    **else**

      (a): **take** the top $M$ items from the probabilistic semantic lexicon entries where $q_i^S$ occurs (e.g., take the top $M$ ranked probability scores $P(q_i^S|\cdot)$ if such entries exist);

      (b): $P_{dm}(q_i^S|d_j^T) = (1-\delta_3)\sum_{l=1}^{M} P(q_i^S|w_l^T)P(w_l^T|d_j^T) + \delta_3 P(q_i^S|Ref^S)$ where $P(q_i^S|w_l^T)$ is calculated using eq. (11.7), and $P(w_l^T|d_j^T)$ using eq. (11.3) or eq. (11.6) ;

    II: **calculate** the probability $P_{muptm}(q_i^S|d_j^T)$:

$$P_{muptm}(q_i^S|d_j^T) = (1-\delta_1)\sum_{k=1}^{K} \overbrace{P(q_i^S|z_k)}^{\text{in source}} \underbrace{P(z_k|d_j^T)}_{\text{in target}} + \delta_1 P(q_i^S|Ref^S)$$

    III: **combine** the calculated probabilities:

    $P(q_i^S|d_j^T) = \lambda P_{dm}(q_i^S|d_j^T) + (1-\lambda)P_{muptm}(q_i^S|d_j^T)$

  IV: **compute** the final query likelihood for the entire query:

$$P(Q^S|d_j^T) = \prod_{i=1}^{m} P(q_i^S|d_j^T) = \prod_{i=1}^{m}\left((1-\delta_1)\sum_{k=1}^{K}\phi_{k,i}\,\theta_{j,k}^T + \delta_1 P(q_i^S|Ref^S)\right)$$

4: **rank** all documents $d_j^T \in \mathcal{DC}^T$ according to their respective scores $P(Q^S|d_j^T)$: $DRank(Q^S)$ ;

**Output**: $DRank(Q^S) \rightarrow$ the ranking of all documents from $\mathcal{DC}^T$ according to their relevance to $Q^S$ ;

### 11.4.3 Test Collections and Queries

We have carried out two conceptually different sets of experiments related to two different evaluation tasks introduced in sect. 11.4.1 to evaluate our retrieval models. The first set of experiments tests the performance of our retrieval models on a less difficult task of the known-item search, where a subset of training documents is used for testing. Another set of experiments has been conducted on target collections that were not used for training beforehand. Here, we deal with a more complex problem, since we want to retrieve documents from a monolingual collection, which might be completely topically unrelated to our training collections (e.g., we train a multilingual topic model on Wikipedia articles and Europarl documents, infer the model on a newswire corpus, and then use the MuPTM-based retrieval models on that newswire corpus). Despite the obvious topical disparity, we believe that by having enough training data from a general domain such as Wikipedia which covers a wide variety of different themes, we are able to learn per-topic word document distributions and infer per-document topic distributions that lead to quality CLIR models, even for topically less general monolingual corpora.

**Wikipedia as a Test Collection for Known-Item Search.** Being document-aligned, Wikipedia data might serve as a framework for the initial evaluation of our models in the less difficult setting, where test articles have already been observed during the MuPTM training. The idea is to simulate the *cross-lingual known-item search*, since it provides a precise semantics and thus removes potential issues with defining an exact information need and assigning relevance judgements. The goal of the proposed cross-lingual known-item search is to find a correct Wikipedia article in the target language $L_T$ with a query provided in the source language $L_S$. The known-item search assumes that only one document is relevant for a specific query.

Since there is no ground truth nor existing queries for this task conducted on this dataset in particular, we have decided to construct the ground truth and the query set by adapting the approach from [13] to the cross-lingual setting. Their approach has already proven useful for the automatic generation of queries for a monolingual known-item search. In the first step, we randomly sample 101 English-Dutch Wikipedia article pairs from our training collection. The target language articles in these article pairs are to be regarded as known items that we want to retrieve. Following that, we generate a *known-item query* by selecting a document pair and constructing a query based on the source language article in the pair. For example, if we have a Wikipedia article pair $A_j = (A_j^S, A_j^T)$, where $A_j^S$ denotes the English article and $A_j^T$ its aligned Dutch counterpart, we are able to generate a known-item query in English from the article $A_j^S$, and then use it to retrieve the article in the target language that is relevant to that

query. The article relevant to the query is then implicitly $A_j^T$ (i.e., the actual known-item). In order to produce the automatic known-item queries along with the ground truth relevance assessments, we have followed these steps:

1. **Pick** a Dutch article $A_j^T$ for which an English query $Q^S$ will be generated.
2. **Initialize** an empty English query $Q^S = \{\}$ for the current article $A_j^T$. Query words are extracted from the article $A_j^S$, and the whole set of English articles is labeled $Coll^S$.
3. **Choose** the query length $len$ with probability $P(len)$. The query length is drawn from a Poisson distribution, with the mean set to the integer closest to the average length of a query for that language from the CLEF collections in order to construct queries of similar length (it was 6 both for English and for Dutch). However, since the query length is drawn from the Poisson distribution, English and Dutch queries for the same article pair are not necessarily of the same length and quality.
4. **For** each word $w_i^S$ in the article $A_j^S$, calculate the probability $P(w_i^S|A_j^S)$, the probability that the word will be sampled from the document model of the article $A_j^S$. Formally, $P(w_i^S|A_j^S)$ is again a mixture between sampling from the article itself and from the entire collection of articles (given in the source language) $Coll^S$ as given by the following formula:

$$P(w_i^S|A_j^S) = (1 - \delta_4)P(w_i^S|A_j^S) + \delta_4 P(w_i^S|Coll^S) \tag{11.9}$$

The quality of the query is influenced by the $\delta_4$ parameter which models noise in the sampling process. As $\delta_4$ decreases to zero, the user is able to recall the content of the article in its entirety. Following the same line of thinking, as $\delta_4$ increases to 1, the user knows that the article exists in the collection, but is not able to recollect any of the words and content relevant to the article. According to [13], setting $\delta_4 = 0.2$ reflects the average amount of noise within the queries for standard IR test collections, so we fix the parameter value to 0.2.

In order to define $P(w_i^S|A_j^S)$, the maximum likelihood estimate of selecting the word $w_i^S$ from the article $A_j^S$, we have opted for the *Popular + Discrimination Selection* strategy from [13] which tries to compromise between *popular words* in a document (i.e., we assume that the user tends to use more frequent words as query words) and *discriminative words* for a document (i.e., the user considers information outside the scope of a document, and tries to construct a query using such query words that discriminate the particular document from the rest of the collection). The

strategy is summarized by the following formula:

$$P(w_i^S|A_i^S) = \frac{tf_{A_j^S}(w_i^S) \cdot \log \frac{|Coll^S|}{df(w_i^S)}}{\sum_{w_l^S \in A_j^S} \left( tf_{A_j^S}(w_l^S) \cdot \log \frac{|Coll^S|}{df(w_l^S)} \right)} \tag{11.10}$$

where $|Coll^S|$ denotes the number of source language articles in the entire collection $Coll^S$, $tf_{A_j^S}(w_i^S)$ denotes the number of occurrences of $w_i^S$ in the article $A_j^S$, and $df(w_j^S)$ is the document frequency of $w_j^S$ in $Coll^S$ (i.e., it measures the number of articles from $Coll^S$ in which $w_j^S$ occurs at least once).

5. **Rank** all words from the article $A_j^S$ based on the scores obtained after employing the previous two equations.
6. **Take** top *len* words from the ranked list as the query words of the known-item query for the article $A_j^S$. At the same time, we have constructed the known-item query for the cross-lingual retrieval of the target language article $A_j^T$ which is aligned to $A_j^S$.

We perform this automatic query generation for 101 article pairs in both directions. The article pairs were randomly sampled from our training Wikipedia data. We design 101 Dutch queries to retrieve English articles and vice versa. For instance, for a Dutch article discussing *halfwaardebreedte (full width at half maximum)*, a query in English is $Q^S = \{width, hyperbolic, variable, deviation\}$.

**CLEF Test Collections.** Another set of experiments tests the ability of proposed CLIR models to perform an ad-hoc subject search in the cross-lingual setting, that is, to rank documents related to the user's information need from a large-scale target collection which covers a broad spectrum of different themes. Therefore, our experiments have been carried out on three standard CLIR test datasets taken from the CLEF 2001-2003 CLIR Evaluation campaigns: the LA Times 1994 (LAT), the LA Times 1994 and the Glasgow Herald 1995 (LAT+GH) in English, and the NRC Handelsblad 94-95 and the Algemeen Dagblad 94-95 (NC+AD) in Dutch. The test datasets are newswire corpora covering a time span of two years and a variety of themes such as politics, sports, art, fashion, geography, literature, etc.

Queries have been extracted from the *title* and *description* fields of all CLEF subjects or each year, as it is standard practice with CLEF test data [237, 238].[3] Queries without relevant documents have been removed from the query sets. Statistics of the test collections are provided in tab. 11.2. Tab. 11.3 shows statistics of the queries used for the test collections.

[3]In order to avoid confusion, we use the term "topics" when referring to the latent concepts obtained by a probabilistic topic model, and "subjects" when referring to the CLEF Evaluation data structure.

### 11.4.4 Training Setup and Parameters of CLIR Models

**Multilingual Topic Model.** We have again used the BiLDA model in all experiments with hyper-parameters $\alpha$ and $\beta$ for the BiLDA training set to the standard symmetric values $50/K$ and 0.01 respectively [278]. We have trained the BiLDA model with different number of topics (400, 1,000 and 2,200) on the combined Wiki+EP corpus. Additionally, for the purpose of comparing retrieval performance when the BiLDA model is trained on different corpora, we have also trained the BiLDA model with $K = 1000$ topics[4] on two different subsets of training corpora: (1) the parallel Europarl sub-corpus (EP)[5], and (2) the comparable Wikipedia corpus (Wiki).

Parameters. The parameter of the Dirichlet prior in the Dirichlet smoothing, $\mu$ is set to the standard value of 1,000 in all models where it is used [323, 330]. Parameters $\delta_1, \delta_2$ and $\delta_3$ are all set to negligible near-zero values. These parameters contribute to the theoretical soundness of the retrieval models, but, due to computational complexity, we do not use counts over a large monolingual reference collection. We use a simple heuristic and assign a small-value constant in all our models instead, since we have empirically detected that these $\delta$

[4]Results with 400 and 2200 topics are comparable and lead to the same conclusions as with $K = 1000$.

[5]As mentioned before, we never exploit the fact that Europarl is sentence-aligned. We use only knowledge of document alignments and nothing else beyond that.

| Collection | Contents | # of Documents |
|---|---|---|
| **LAT** | LA Times 94 (EN) | 110,861 |
| **LAT+GH** | LA Times 94 (EN)<br>Glasgow Herald 95 (EN) | 166,753 |
| **NC+AD** | NRC Handelsblad 94-95 (NL)<br>Algemeen Dagblad 94-95 (NL) | 190,604 |

Table 11.2: Statistics of the CLEF 2001-2003 CLIR test collections.

| CLEF Subjects (Year: Topic Nr.) | # of Queries (with Rel. Docs) | Used for Collections |
|---|---|---|
| NL 2001: 41-90 | 47 | LAT |
| NL 2002: 91-140 | 42 | LAT |
| NL 2003: 141-200 | 53 | LAT+GH |
| EN 2001: 41-90 | 50 | NC+AD |
| EN 2002: 91-140 | 50 | NC+AD |
| EN 2003: 141-200 | 56 | NC+AD |

Table 11.3: Statistics of used queries (CLEF test collections).

parameters do not have any significant impact on the results. The interpolation parameter in the Jelinek-Mercer smoothing is set to $\lambda = 0.3$ in all experiments where it is used, which assigns more weight to the topical representation of documents. We have also tried to experiment with different values of $\lambda$ (e.g., $\lambda = 0.7$, which assigns more weight to the DM representation of documents), and the results slighly differ in absolute numbers, but all key conclusions remain the same. Therefore, we present only the results obtained by setting $\lambda = 0.3$.

We have also empirically detected that the optimal value for $M$ is 10, so we have used top 10 items from the ranked list for each probabilistic semantic lexicon entry in all experiments with the SemLex-Basic and the MuPTM-SemLex models.[6] Since the research reported in this chapter was conducted in the earlier stage of this thesis, all probabilistic semantic lexicons were obtained from Wiki+EP by then best-scoring TI+Cue model of cross-lingual semantic similarity [312] (see also chapter 6) with the number of topics $K = 2200$, and all reported results utilize that probabilistic lexicon.

### 11.4.5 Evaluation Metrics

Since the goal of the known-item search for a Wikipedia article is to find and retrieve a single article relevant to a query, we again report results in the form of the *Top M* accuracy ($Acc_M$). We report $Acc_1$ (the percentage of queries for which the only relevant document is retrieved as the first candidate in the ranked list of articles) and $Acc_5$ (the percentage of queries for which the only relevant document is retrieved among the top 5 retrieved documents in the list).

The main evaluation measure in all experiments conducted on the CLEF test collections is the standard IR measure of *mean average precision* (MAP) [185]. Given a set of source language queries $\mathcal{Q}^S$ that consists of $|\mathcal{Q}^S|$ queries, the MAP score is calculated as follows:

$$MAP(\mathcal{Q}^S) = \frac{\sum_{Q^S \in \mathcal{Q}^S} AP(Q^S)}{|\mathcal{Q}^S|} \tag{11.11}$$

Here, $AP(Q^S)$ is the average precision score for one query $Q^S$ from the query set, and it is calculated as follows:

$$AP(Q^S) = \frac{\sum_{l=1}^{|DRank(Q^S)|} Precision(l) \cdot rel(l)}{|REL(Q^S)|} \tag{11.12}$$

[6] We have experimented with different values, $M = 1, 3, 5, 10, 20$, and have empirically detected that $M = 10$ displays the best results overall, although variations when using other values for $M$ are in most cases minimal.

$DRank(Q^S)$ is a ranked list of documents sorted according to their relevance to the query $Q^S$. $Precision(l)$ denotes the precision score for $Q^S$ at cut-off $l$ in the ranked list of documents from the target collection, while $rel(l)$ is an indicator function equaling 1 if the document at rank $l$ in the list is a relevant document according to ground truth relevance assessments, zero otherwise. $|REL(Q^S|$ denotes the number of documents relevant to the query $Q^S$ given by ground truth relevance assessments. In some experiments, as an extra evaluation metric, we also provide *11-pt recall-precision curves*.

## 11.5 Experiments, Results and Discussion

### 11.5.1 A Short Overview

This section reports our experimental results for the two main tasks introduced in the previous section. We test our retrieval models from sect. 11.3 in the task of the known-item search for Wikipedia articles and report our findings. As the next step, we carry out different experiments for English-Dutch and Dutch-English cross-lingual information retrieval on CLEF test collections:

(1) We compare the MuPTM-Basic against several baselines that have also tried to exploit latent topic spaces for CLIR: (i) cross-lingual latent semantic analysis (CLSA) trained on concatenated paired documents [81], (ii) standard monolingual LDA trained on concatenated paired documents [261]. We also compare MuPTM-Basic with the DM-Basic model from sect. 11.3.2. We want to prove the soundness and the utility of the MuPTM-Basic model and, consequently, other models that are later built upon the foundation established by the MuPTM-Basic model (MuPTM-DM and MuPTM-SemLex).

(2) We provide an extensive evaluation over all CLEF test collections with all MuPTM-based models (MuPTM-Basic, MuPTM-DM and MuPTM-SemLex).

(3) We compare our MuPTM-based models with similar models tailored for monolingual retrieval (queries and documents given in the same language) and a model that uses the *Google Translate* tool to translate queries and then performs monolingual retrieval, and measure the decrease of performance for CLIR.

(4) We also compare the combined MuPTM-SemLex model with the SemLex-Basic model that uses only evidence of the shared words and knowledge from the MuPM-based probabilistic semantic lexicon.

(5) We compare results for all test collections when the multilingual topic model is trained on different types of training data (parallel, comparable and combined)

and show that including comparable data boosts retrieval performance.

(6) We report a mismatch between perplexity scores in a quantitative intrinsic evaluation and ex vivo extrinsic evaluation, that is, final retrieval scores.

### 11.5.2 Experiment 0: Cross-Lingual Known-Item Search for Wikipedia Articles

**Experimental Setup and Results.** The cross-lingual known-item search has been carried out for 101 pairs of Wikipedia articles randomly sampled from 7,612 pairs constituting the English-Dutch Wikipedia training corpus. Experiments have been conducted for both possible retrieval directions (English to Dutch and Dutch to English). The BiLDA model was beforehand trained on the Wiki+EP corpus. To make the search for the single relevant document for each query even more difficult, we have also included the Europarl documents in the search space. Our search space then consists of 13,818 documents from all training document pairs. Results for the DM-Basic model and the SemLex-Basic model in terms of $Acc_1$ and $Acc_5$ scores are provided in tab. 11.4. Results for MuPTM-Basic, MuPTM-Uni and MuPTM-SemLex both search directions are provided in tab. 11.5 and tab. 11.6.

| | EN Queries, NL Documents | | NL Queries, EN Documents | |
|---|---|---|---|---|
| | DM-Basic | SemLex-Basic | DM-Basic | SemLex-Basic |
| $Acc_1$ | 0.406 | 0.520 | 0.485 | 0.630 |
| $Acc_5$ | 0.525 | 0.610 | 0.584 | 0.723 |

Table 11.4: $Acc_1$ and $Acc_5$ scores in both search directions for DM-Basic and SemLex-Basic in the cross-lingual known-item search for Wikipedia articles.

**Discussion.** We have drawn several conclusions based on the obtained results:

(i) Tab. 11.4 reveals that using the knowledge from probabilistic semantic lexicon entries significantly helps in improving overall search performance. However, these results are still much lower than results obtained by combining shared words and the knowledge from lexicon entries with the "MuPTM-based" part from the MuPTM-Basic model (compare results in tab. 11.5 and tab. 11.6).

(ii) The MuPTM-Basic model is outperformed by the MuPTM-DM and the MuPTM-SemLex model which exploit more different evidences and try to use them in the LM IR modeling framework. We conclude that the combination of cross-lingual evidences leads to better retrieval and search models, even when

| $K$ | $Acc_1$ | | | $Acc_5$ | | |
|---|---|---|---|---|---|---|
| | MuPTM-Basic | MuPTM-DM | MuPTM-SemLex | MuPTM-Basic | MuPTM-DM | MuPTM-SemLex |
| 400 | 0.198 | 0.673 | 0.668 | 0.396 | 0.792 | 0.792 |
| 1000 | 0.406 | 0.724 | 0.747 | 0.674 | 0.822 | 0.852 |
| 2200 | 0.465 | 0.757 | **0.787** | 0.773 | 0.901 | **0.941** |

Table 11.5: $Acc_1$ and $Acc_5$ scores for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex in the cross-lingual known-item search. EN queries, NL target articles.

| $K$ | $Acc_1$ | | | $Acc_5$ | | |
|---|---|---|---|---|---|---|
| | MuPTM-Basic | MuPTM-DM | MuPTM-SemLex | MuPTM-Basic | MuPTM-DM | MuPTM-SemLex |
| 400 | 0.376 | 0.782 | 0.770 | 0.604 | 0.862 | 0.881 |
| 1000 | 0.535 | 0.780 | 0.770 | 0.792 | 0.921 | 0.931 |
| 2200 | 0.594 | 0.841 | **0.841** | 0.862 | 0.961 | **0.980** |

Table 11.6: $Acc_1$ and $Acc_5$ scores for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex in the cross-lingual known-item search. NL queries, EN target articles.

the evidences are not completely disjunct. That conclusion will be more firmly supported by later experiments on the CLEF collections.

(iii) MuPTM-DM and MuPTM-SemLex display comparable results, with a slight advantage for the MuPTM-SemLex model. The observation is explained if we investigate the structure of the queries. Many Wikipedia articles describe people, toponyms or specific concepts where many words are shared between the Dutch and English vocabularies. In such a setting, a lexicon helps to a lesser extent.

(iv) We have successfully applied a method from [13] to automatically generate queries for known-item search and we have adapted it to the cross-lingual setting. Moreover, Azzopardi et al. [13] assert that their method still suffers from the insufficient *replicative validity* and *predictive validity* (i.e., an automatically generated query should really behave as a query generated from the user, and retrieved articles should be similar in both cases). Using a thorough evaluation, they claim that automatically generated queries typically lead to lower retrieval scores, which leads to conclusion that the results with *real-life* manual queries might be even higher than presented in the tables. We have tested the models with manually generated queries, and the results are indeed comparable to the ones presented, and the same general conclusions may be drawn.

### 11.5.3 Experiment I: Comparison of the MuPTM-Basic Model with Baseline Models

**Results and Discussion.** From now on, all experiments are conducted on the CLEF test collections.

The MuPTM-Basic model serves as the backbone of the two more advanced MuPTM-based retrieval models (MuPTM-DM and MuPTM-SemLex). Since we want to make sure that the MuPTM-Basic model constructs a firm and sound language pair independent foundation for building more complex cross-lingual retrieval models, we compare it to state-of-the-art systems which also aim to build a CLIR model based on the idea of shared latent concept spaces: (a) the cross-lingual latent semantic analysis (CLSA) as described in [81, 41], which constructs a reduced (latent) vector space trained on concatenated paired documents in two languages, and (b) the standard LDA model trained on the merged document pairs [261]. We have trained the CLSA model and the standard LDA model on the combined Wiki+EP corpus with 400 and 1000 dimensions (concepts/topics) and have compared the retrieval scores with the scores obtained by our MuPTM-Basic model that relies the BiLDA model with the same number of topics. The MuPTM-Basic model outscores the other two models by a huge margin. The MAP scores for CLSA and standard LDA are similar and very low, and vary between the MAP score of 0.01 and 0.03 for all experimental runs, which is significantly worse than the results of the MuPTM-Basic model. The MAP scores for the MuPTM-Basic model for NL 2001, NL 2002, and NL 2003 with $K = 1000$ are 0.197, 0.140 and 0.123 respectively, while the MAP scores for EN 2001, EN 2002, and EN 2003 for K=1000 are 0.145, 0.137, and 0.171, respectively (see tab. 11.7).

One reason for such a huge difference in scores might be the ability to infer the BiLDA model on a new test collection (due to its fully generative semantics) more accurately. CLSA for CLIR reported in the literature always uses the same corpus (or subsets of the same corpus) for training and testing, while this setting requires inferring the model on a test corpus which is not by any means content-related to the training corpus. BiLDA has a better statistical foundation by defining the common per-document topic distribution $\theta$, which allows inference on new documents based on the previously trained model and also avoids the problem of overfitting inherent to the CLSA model. Another problem with the baseline methods lies in the concatenation of document pairs, since one language might dominate the merged document. On the other hand, BiLDA keeps the structure of the original document space intact.

The MAP scores of the DM-Basic model for NL 2001, NL 2002, and NL 2003 are 0.027, 0.034 and 0.029 respectively, while the MAP scores for EN 2001, EN

2002, and EN 2003 are 0.064, 0.103 and 0.083 respectively. Comparison of the 11-pt recall-precision curves for the MuPTM-Basic model and the DM-Basic model are presented in fig. 11.5a and fig. 11.5b. All these results justify the use of a multilingual topic model (e.g., BiLDA) in other more complex retrieval models.

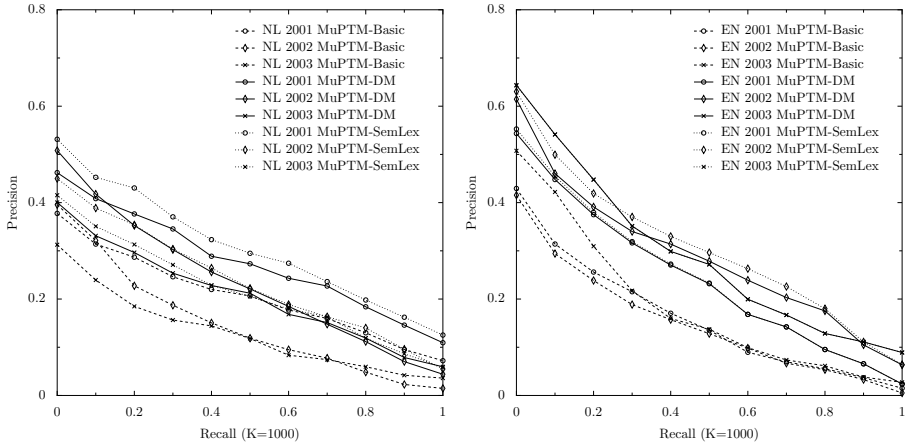
### 11.5.4 Experiment II: Comparison of MuPTM-Basic, MuPTM-DM and MuPTM-SemLex

In this section, we aim to compare the three retrieval models that rely on the usage of MuPTM-based per-document topic distributions, once the multilingual topic model is inferred on a target collection, that is, our goal is to test whether the knowledge from shared words (as in the MuPTM-DM model), and the knowledge of the shared words combined with the knowledge from MuPTM-based probabilistic semantic lexicons (as in the MuPTM-SemLex model) positively affect retrieval quality.

**Comparison of Models with a Fixed Number of Topics** ($K = 1000$)**.** The MuPTM-Basic model, the MuPTM-DM model and the MuPTM-SemLex model have been evaluated on all CLEF test collections, with the number of topics initially fixed to $1,000$. Fig. 11.3a shows the 11-pt recall precision curves obtained by applying all three models to EN test collections with NL queries, while fig. 11.3b shows the curves for NL test collections and EN queries.

As the corresponding figures show, the MuPTM-Basic model seems to be too coarse to be used as the only component of an IR model (e.g., due to its limited number of topics, words in queries unobserved during training). In other words, a sole document representation by means of per-document topic distributions is not sufficient to produce quality retrieval models. However, combining the topical representation with words shared across languages and lexicon entries from MuPTM-induced lexicons leads to a drastic increase in results. Results of the MuPTM-SemLex model which scores better than the MuPTM-DM model are especially encouraging. The MuPTM-DM relies solely on shared words, which clearly makes it language pair biased, since its performance heavily relies on the amount of shared words (or the "degree of vicinity" between the two languages involved in the retrieval process). On the other hand, the MuPTM-SemLex model has been envisioned for CLIR between distant language pairs (e.g., a similar model for the English-Bengali CLIR is investigated in [98]).

**Varying the Number of Topics.** In the next set of experiments we test the importance of the MuPTM-based probabilistic semantic lexicon, and the behavior of the two best-scoring models when we vary the number of




(a) NL queries, EN test collections. (b) EN queries, NL test collections.

Figure 11.3: 11-pt recall-precision curves for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex for both retrieval directions. Multilingual topic model is BiLDA, $K = 1000$. Training corpus is Wiki+EP.

topics during the BiLDA training. We have carried out experiments with the CLIR models relying on BiLDA trained with different numbers of topics ($K = 400, 1000, 2200$). The MAP scores of MuPTM-Basic, MuPTM-DM and MuPTM-SemLex for all campaigns are presented in tab. 11.7, while fig. 11.4 shows the associated recall-precision curves of the two best-scoring CLIR models.

| Queries/$K$ | MuPTM-Basic | | | MuPTM-DM | | | MuPTM-SemLex | | |
|---|---|---|---|---|---|---|---|---|---|
| | 400 | 1000 | 2200 | 400 | 1000 | 2200 | 400 | 1000 | 2200 |
| NL 2001 | 0.178 | 0.197 | 0.203 | 0.233 | 0.267 | 0.281 | **0.300** | 0.294 | 0.297 |
| NL 2002 | 0.112 | 0.140 | 0.137 | 0.209 | 0.225 | 0.221 | **0.242** | 0.226 | 0.224 |
| NL 2003 | 0.078 | 0.123 | 0.078 | 0.161 | 0.199 | 0.166 | 0.206 | **0.208** | 0.181 |
| EN 2001 | 0.127 | 0.145 | 0.162 | 0.220 | 0.228 | 0.240 | 0.229 | 0.237 | **0.243** |
| EN 2002 | 0.093 | 0.137 | 0.141 | 0.246 | 0.268 | 0.267 | 0.271 | **0.287** | 0.278 |
| EN 2003 | 0.098 | 0.171 | 0.153 | 0.239 | 0.278 | 0.245 | 0.239 | **0.278** | 0.250 |

Table 11.7: MAP scores on all CLEF test collections for MuPTM-Basic, MuPTM-DM and MuPTM-SemLex, where BiLDA was trained with different number of topics (400, 1000, 2200). Training corpus is Wiki+EP.

**Discussion.** We observe several interesting phenomena based on the results from tab. 11.7 and fig. 11.4:

(i) The MuPTM-SemLex model obtains the best scores for all test collections

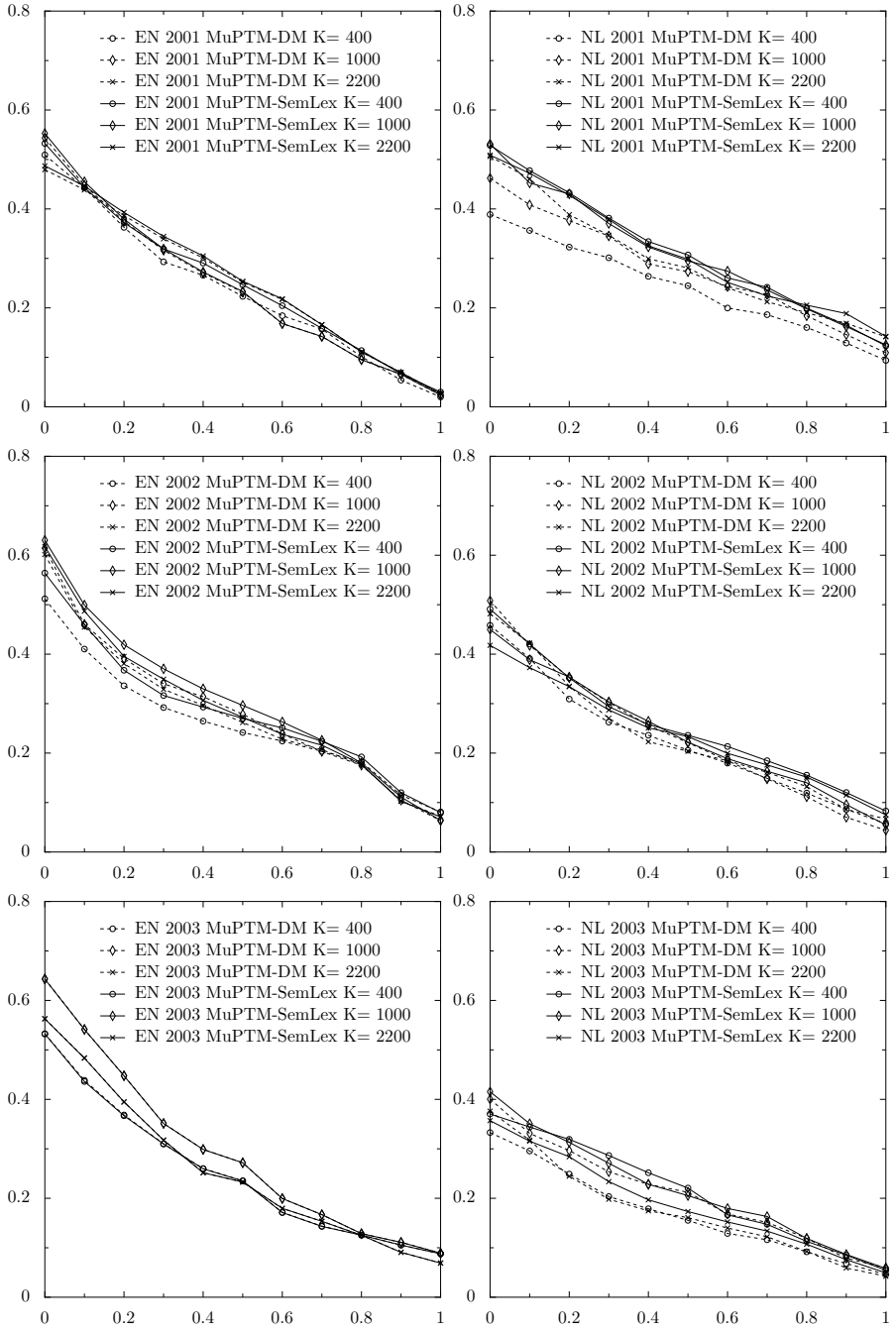
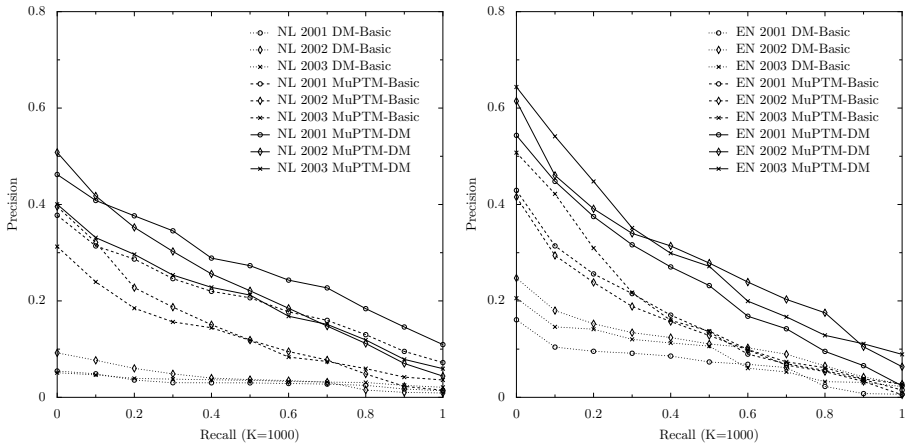



Figure 11.4: 11-pt recall-precision curves for MuPTM-DM and MuPTM-SemLex for all CLEF test collections. Multilingual topic model is BiLDA. Training corpus is Wiki+EP.

which proves the intuition that additional evidences coming from a MuPTM-based probabilistic semantic lexicon will improve the retrieval scores and lead to a better model.

(ii) The margins between scores of the MuPTM-DM and the MuPTM-SemLex model are generally higher for campaigns with Dutch queries. The reason why the Dutch-English lexicon seems to be more helpful might lie in the fact that much more English words are observed in our Dutch vocabulary than vice versa. If that is the case, than the knowledge from the lexicon is used less frequently, and the MuPTM-SemLex model relies more on shared words, which brings it closer to the MuPTM-DM model. On the other hand, less Dutch words are observed in the English vocabulary, and one needs to turn to the evidences from the semantic lexicon more often. In order to support this intuition which explains the results from fig. 11.4, we have computed the average percentage of shared words in both English and Dutch queries. The average percentage of shared words is 55.6% per English query, and only 18.9% per Dutch query. This difference in percentage of shared words comes mostly from the English terms such as named entities that are often used in parallel with Dutch terms in Dutch news texts. For instance, when a Dutch news article discusses the London or the New York Stock Exchange, it often uses the exact English term, while an English article, of course, will not include the Dutch translation.

(iii) Due to a high percentage of shared words, especially in English queries (see the previous discussion item), it may be possible that the MuPTM-DM model




(a) NL queries, EN test collections. (b) EN queries, NL test collections.

Figure 11.5: Comparison of DM-Basic, MuPTM-Basic, and MuPTM-DM as their combination. BiLDA with $K = 1000$. Training corpus is Wiki+EP.

draws its performance mainly from the part specified by the DM-Basic model. However, as presented in fig. 11.5a and fig. 11.5b, that possibility has been denied, and the final combined MuPTM-DM model clearly works as a positive synergy between the two simpler basic models, outperforming both of them. However, MuPTM-Basic provides higher scores than DM-Basic and is therefore more important for the overall performance of the combined model.

(iv) The margins between scores of the MuPTM-DM and the MuPTM-SemLex model are generally higher for the lower number of topics in campaigns with Dutch queries, where the semantic lexicons are used more extensively. With less topics, per-topic word distributions and per-document topic distributions are too coarse, so more cross-language evidence comes from the lexicon itself. By increasing the number of topics, these distributions become more fine-grained, and more and more evidence that initially came from the lexicon, is now captured by the "MuPTM-part" of the MuPTM-SemLex model. We have encountered overlaps of the evidences which lead to similar retrieval scores. The scores obtained by the MuPTM-SemLex model are still higher than the scores of MuPTM-DM, since some of the evidences can be found only in the lexicon, regardless of the number of topics.

(v) Although a larger number of topics should intuitively lead to a more fine-grained model with better estimated per-topic word distributions and per-document topic distributions and, consequently, to a better retrieval model, this is clearly not the case. If we set the number of topics to a value too high, the topics will become less informative and descriptive as the evidences tend to disperse over all topics. In the most extreme case, each word could be a topic on its own, but how informative is that? And what can we learn from it? One of the main disadvantages of the BiLDA model is the need to define the number of topics before its training takes place. It does not have the ability to dynamically redefine the number of topics to adjust to the training data in an optimal way. Therefore, one line of future research lies in designing non-parametric multilingual probabilistic topic models similar to the non-parametric topic models in the monolingual setting [26, 176, 201], that need not set the number of topics in advance, but are able to adapt to training data. However, in our preliminary studies with non-parametrized multilingual topic models [68], we have not detected any improvement in the overall retrieval results. These topic models give rise to other research questions and challenges that fall out of the scope of this thesis.

(vi) Our best scoring MuPTM-SemLex model is competitive with models which utilize external translation dictionaries or MT systems. For instance, the model would have been placed among the top 5 retrieval systems for the CLEF 2002 Bilingual to Dutch task. It would have been placed among the top 3 retrieval systems for the CLEF 2001 Bilingual to Dutch task, and outperforms the only

participating system in the CLEF 2002 Dutch to English task (MAP: 0.1495) [237, 238]. All these systems construct queries from *title* and *description* or *title*, *description* and *narrative* fields from the CLEF subjects in the same manner as done here, but they use translation dictionaries manually constructed or obtained from sentence-aligned parallel data.

### 11.5.5 Experiment III: Comparison with Monolingual MoPTM-Based Models and MuPTM-Based Models that Use an External Translation Resource

**Motivation and Results.** With this set of experiments, we investigate the effectiveness of our MuPTM-based CLIR models. Therefore, we compare the three MuPTM-based models evaluated in sect. 11.5.4 with another four models: (1) a model that performs monolingual retrieval in the same fashion as our CLIR MuPTM-Basic model (*MoPTM-Basic*), (2) a model that performs monolingual retrieval in the same fashion as our CLIR MuPTM-DM model (*MoPTM-DM*), (3) a model that uses *Google Translate* to perform query translation, and then performs monolingual retrieval using MoPTM-Basic (*GT+MoPTM-Basic*), (4) a model that uses *Google Translate* in the same manner, and then employs the monolingual MoPTM-DM model (*GT+MoPTM-DM*). In order to use these models, we have trained standard monolingual LDA with $K = 1000$ topics for both English and Dutch side of our training corpora. MAP scores for these models are presented in tab. 11.8, while MAP scores for our MuPTM-based CLIR models have already been presented in tab. 11.7. MoPTM-Basic and MoPTM-DM operate with queries given in the same language as as the target document collection. To remain consistent throughout the text and facilitate the comparison of different models, we do not change the naming conventions for the queries and document collections in tab. 11.8. However, the reader has to be aware that when applying MoPTM-Basic and MoPTM-DM, NL 2001 actually means - English queries (instead of Dutch queries as in CLIR) to retrieve English documents. We are then allowed to compare results of all the models (both monolingual and CLIR) for the NL 2001 campaign.

| Model/Queries | NL 2001 | NL 2002 | NL 2003 | EN 2001 | EN 2002 | EN 2003 |
|---|---|---|---|---|---|---|
| MoPTM-Basic | 0.280 | 0.216 | 0.241 | 0.132 | 0.143 | 0.130 |
| MoPTM-DM | 0.399 | 0.336 | 0.379 | 0.260 | 0.289 | 0.326 |
| GT+MoPTM-Basic | 0.186 | 0.185 | 0.226 | 0.125 | 0.115 | 0.116 |
| GT+MoPTM-DM | 0.307 | 0.275 | 0.348 | 0.230 | 0.240 | 0.244 |

Table 11.8: MAP scores on all CLEF test collections for MoPTM-Basic, MoPTM-DM, GT+MoPTM-Basic and GT+MoPTM-DM. Standard monolingual LDA trained on monolingual English and Dutch data. Wiki+EP. $K = 1000$.

**Discussion.** By examining the results in tab. 11.7 and tab. 11.8, we derive several conclusions:

(i) As expected, the monolingual MoPTM-DM model that combines two different document representations outperforms our CLIR models, although the difference in scores is much more discernible when performing monolingual retrieval in English. We attribute that observation to the quality of our training data. The English side of our Wikipedia data contains more information and articles of a higher quality, which altogether leads to better estimated latent topics, which then consequently leads to better statistical retrieval models. Following the same line of thinking, while MAP scores for the MoPTM-Basic model for Dutch are comparable to the scores of MuPTM-Basic when submitting queries in English, MoPTM-Basic for English scores much better than MuPTM-Basic with Dutch queries.

(ii) Low results for MoPTM-DM for monolingual Dutch retrieval when we train standard LDA on monolingual data also refer to the fact that the Dutch side of our training corpus is of a lesser quality.

(iii) A significant downtrend in performance for both retrieval directions has been observed when we use *Google Translate* to perform query translation and then perform monolingual retrieval. *Google Translate* might also introduce some errors in the translation process. That conclusion underpins the conclusions drawn by [79].

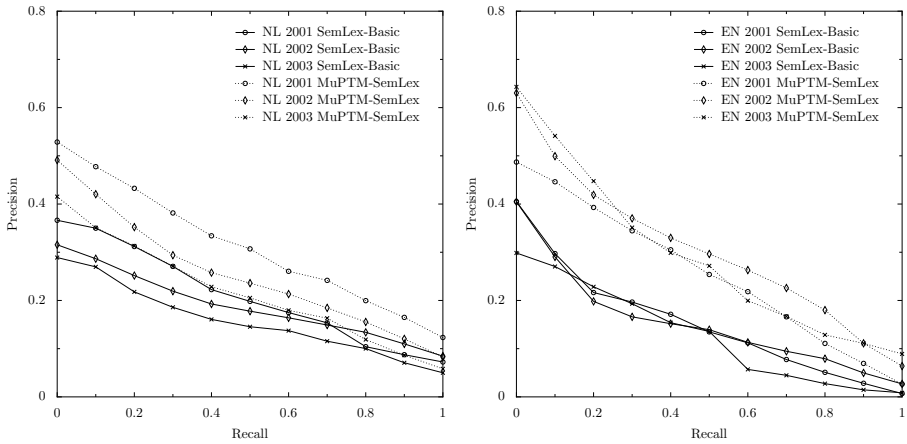
(iv) Our combined evidence-rich CLIR models outperform GT+MoPTM-DM for English queries and Dutch text collections. One of the reasons for that phenomenon might again be the errors in the translation process introduced by *Google Translate*. Moreover, many words from English queries are shared across languages and therefore also present in Dutch documents, and our MuPTM-DM and MuPTM-SemLex models are able to capture that tendency.

(v) For almost all CLEF campaigns, our MuPTM-DM and MuPTM-SemLex models display performance that is comparable with or even better than performance of the GT+MoPTM-DM model, a model that uses an external translation resource to directly translate queries. Our models thus become extremely important for language pairs where such a translation system or a dictionary is low-quality or unavailable.

### 11.5.6 Experiment IV: Comparison of MuPTM-SemLex and SemLex-Basic

**Motivation and Results.** We also want to compare our best scoring MuPTM-SemLex model with the SemLex-Basic model. While MuPTM-SemLex blends evidences from MuPTM-induced semantic lexicons and shared words with evidences from the output probability distributions of a multilingual topic model, SemLex-Basic utilizes only the knowledge of shared words and the knowledge coming from probabilistic semantic lexicons. We have already proven that the combined, evidence-rich models yield better scores than the MuPTM-Basic model that exploits only evidences in the form of per-topic word and per-document topic distributions. We now aim to prove that the evidence-rich MuPTM-SemLex model also scores better than the more straightforward SemLex-Basic model that uses only *lexical* evidences. MAP scores for the SemLex-Basic model are 0.1998, 0.1810 and 0.1513 for NL 2001, NL 2002 and NL 2003 (Dutch queries, English documents) respectively, and 0.1412, 0.1378 and 0.1196 for EN 2001, EN 2002 and EN 2003 (English queries, Dutch documents). The best MAP scores for MuPTM-SemLex are given in tab. 11.7. Fig. 11.6a shows the comparison of the associated 11-pt recall-precision diagrams for all English collections (with queries in Dutch), while fig. 11.6b shows the comparison for all Dutch collections (with queries in English).

**Discussion.** Fig. 11.3b and fig. 11.3a have already shown the superiority of the MuPTM-SemLex model over the MuPTM-Basic model. The results in this




(a) NL queries, EN test collections. (b) EN queries, NL test collections.

Figure 11.6: 11-pt recall-precision curves for SemLex-Basic and MuPTM-SemLex. BiLDA. Training corpus is Wiki+EP.

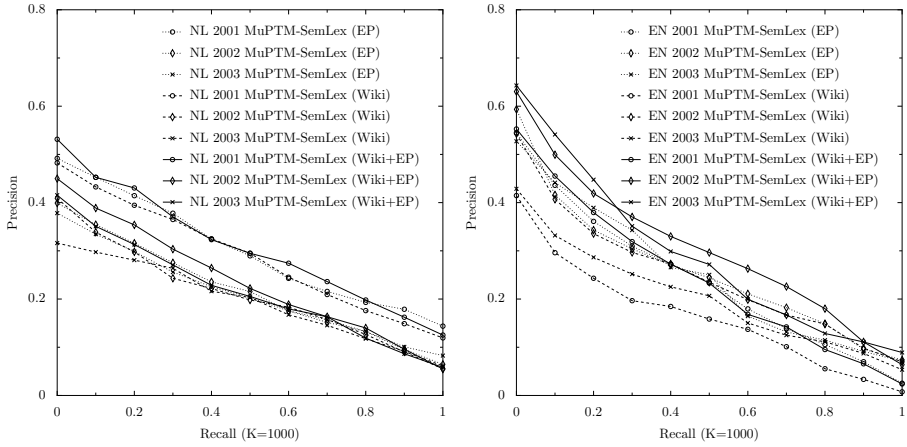
section again show the superiority of the combined MuPTM-SemLex model over the SemLex-Basic model. The fact that the MuPTM-SemLex model combines the evidences from the other two basic models makes it the most powerful model. The other two models utilize only subsets of the available evidences which makes them more error-prone. For instance, if the semantics of a word from a query is not captured by the "MuPTM-part" (as in the MuPTM-Basic model), that model is unable to retrieve any documents strongly related to that word. On the other hand, if the same problem occurs for the combined MuPTM-SemLex model, the model still has a possibility to look up for an aid in the lexicon. Additionally, if a document scores good for more than one evidence, it strengthens the belief that the document might be relevant for a query.

### 11.5.7 Experiment V: Training with Different Types of Corpora

**Motivation and Results.** In the next set of experiments with our CLEF test collections, we measure the performance of the MuPTM-based CLIR models on three different types of corpora (EP, Wiki, their combination: Wiki+EP) with $K = 1000$ topics. We want to find out if and how Wikipedia training data help the retrieval. Moreover, we want to test our *"the more the merrier"* assumption that more training data lead to better estimates of the output probability distributions of a multilingual topic model and, following that, better retrieval models. The best-scoring MuPTM-DM and MuPTM-SemLex models have been used in all experimental runs in this section. Tab. 11.9 displays the MAP scores over all CLEF test runs, while fig. 11.7a shows recall-precision curves for the campaigns with Dutch queries and English documents, and fig. 11.7b for the other retrieval direction.

| Queries | MuPTM-DM | | | MuPTM-SemLex | | |
|---|---|---|---|---|---|---|
| | EP | Wiki | Wiki+EP | EP | Wiki | Wiki+EP |
| NL 2001 | 0.259 | 0.180 | 0.267 | 0.290 | 0.280 | **0.294** |
| NL 2002 | 0.179 | 0.179 | 0.225 | 0.209 | 0.199 | **0.226** |
| NL 2003 | 0.181 | 0.125 | 0.199 | 0.206 | 0.190 | **0.208** |
| EN 2001 | 0.229 | 0.148 | 0.228 | 0.228 | 0.151 | **0.237** |
| EN 2002 | 0.237 | 0.218 | 0.268 | 0.240 | 0.232 | **0.287** |
| EN 2003 | 0.240 | 0.192 | 0.278 | 0.240 | 0.196 | **0.278** |

Table 11.9: MAP scores on all CLEF test collections for MuPTM-DM and MuPTM-SemLex, where BiLDA was trained on different corpora (EP, Wiki, and Wiki+EP). $K = 1000$.




(a) NL queries, EN test collections. (b) EN queries, NL test collections.

Figure 11.7: Comparison of the 11-pt recall-precision curves values for MuPTM-SemLex, where BiLDA was trained on different corpora (EP, Wiki and Wiki+EP). $K = 1000$.

**Discussion.** The results lead us to several conclusions:

(i) They show that the comparable general-domain Wikipedia data may be used to train a multilingual topic model and reasonable CLIR results can still be acquired. For some experiments (EN 2002, NL 2001 and NL 2002), the results with the model trained on the Wikipedia data are comparable to the results with the model trained on the parallel document-aligned Europarl corpus (especially for the MuPTM-SemLex model, when the additional MuPTM-based semantic lexicon knowledge is employed). For these campaigns we observe major improvements when the multilingual topic model is trained on the combined corpus. On the other hand, some experiments where retrieval models rely on the topic model trained solely on Wikipedia have led to much worse scores than the scores of models relying on the topic model trained on Europarl (e.g. NL 2003, EN 2001). For these experiments, we do not observe any major improvement after we enrich our training data with Wikipedia data. However, we believe that extracting more Wikipedia articles as training data might resolve this problem.

(ii) Tab. 11.9 also reveals that accumulating more training data by adding comparable Wikipedia documents to a parallel corpus is definitely not harmful, and in most cases increases the quality of the estimated latent cross-lingual topics, which also leads to better CLIR models.

(iii) These experiments again reveal a clear advantage of using the automatically extracted MuPTM-based semantic lexicons, since the MuPTM-SemLex model,

which uses the lexicon, displays better results than MuPTM-DM over all CLEF test collections. A more thorough analysis and comparison of these two models is provided in sect. 11.5.4.

### 11.5.8 Experiment VI: Perplexity and Retrieval

As already discussed in sect. 4.5, the perplexity of a model measures its adaptivity to a collection of unseen documents, that is, its ability to explain the unseen text collection. In theory, a lower perplexity score implies a better model. Tab. 11.10 shows perplexity scores after the BiLDA models with different number of topics $K$ were inferred on the CLEF test collections. One might observe that the increase of perplexity scores correlates with the increase of the number of topics.

| Test collection | K=400 | K=1000 | K=2200 |
|---|---|---|---|
| LAT | **111.12** | 215.98 | 437.11 |
| LAT+GH | **107.91** | 210.15 | 432.76 |
| NC+AD | **110.85** | 219.45 | 527.43 |

Table 11.10: Perplexity scores after the inference of the BiLDA model (trained on the Wiki+EP corpus) on the CLEF test collections.

A comparison of the results reported in tab. 11.10 and tab. 11.7 clearly indicates that the theoretical measure of perplexity, often used to intrinsically compare the quality of probabilistic topic models, does not guarantee a better performance in real-life applications such as CLIR. The same general conclusion for language models in information retrieval has also been drawn by Azzopardi et al. [14].

## 11.6 Conclusions and Future Work

In this chapter, aiming to address requirement R3 and research question RQ3 from chapter 1, we have proposed and constructed a new probabilistic language-pair independent framework for cross-language information retrieval. The framework is built upon the knowledge coming from multilingual probabilistic topic models trained on non-parallel data. The proposed framework does not utilize any type of an external translation resource such as a machine translation system or a dictionary which is expensively hand-crafted or extracted from parallel data.

This chapter presents and discusses several new probabilistic MuPTM-based CLIR models built within this framework. These models exploit different

evidences in the retrieval process (e.g., while the MuPTM-Basic retrieval model exploits only output distributions coming from a multilingual topic model, the MuPTM-SemLex retrieval model combines that knowledge with the knowledge of shared words and probabilistic lexicon entries). We have thoroughly evaluated and compared all our models using our manually constructed Wikipedia test set for known-item search and standard test collections from the CLEF 2001-2003 CLIR campaigns, presenting and explaining their key advantages and shortcomings. We have shown that our combined models, which fuse more different retrieval clues obtain the best retrieval scores in general.

The importance of the proposed framework lies in the fact that it allows for constructing more elaborate retrieval models which capture additional evidence (for instance, a knowledge from an external lexicon or a knowledge of cognate pairs are easily incorporated into the retrieval process and yield new retrieval models). In addition, the construction of other probabilistic retrieval models that go beyond the query likelihood modeling principle is also possible within the same framework. For instance, more advanced and more robust probabilistic CLIR models relying on the framework of relevance modeling are introduced in the next chapter. Another advantage of the proposed framework is that it provides a unified approach to monolingual IR, CLIR and MIR (again, more discussion on that in the following chapter).

Since the estimation of multilingual probabilistic topic models is done *offline* (following the "*learn once, use many*" principle), there should be no major restrictions in applying once estimated topic models to different target document collections in different domains. Another straightforward extension of the framework is porting it to other language pairs besides the English-Dutch language pair which was used for evaluation in this chapter.

The modularity of the framework allows for investigating various paths of future work. In addition to building new retrieval models which capture additional clues in the retrieval process (see the following chapter), one may again test different multilingual topic models (see sect. 4.6 in chapter 4) or any other model that induces latent cross-lingual concepts (see sect. 4.3 in chapter 4) within the same framework, or apply the IR approach to other tasks besides the core known-item and subject search. For instance, in [342] and [341], we have proposed a new task of linking Pinterest pins to online Web shops, framed it as an IR problem and showed how to apply this very same probabilistic framework in that novel task. Another relevant task where this retrieval framework may be easily applied and further investigated is retrieving answers to user-generated questions from existing repositories of community-driven question-and-answer sites (such as Yahoo! Answers, Quora or Ask.com) [208].

## 11.7 Related Publications

[1] **I. Vulić**, W. De Smet, and M.-F. Moens. "Cross-language information retrieval with latent topic models trained on a comparable corpus," in *Proceedings of the 7th Asian Information Retrieval Societies Conference (AIRS)*, vol. 7097 of *Lecture Notes in Computer Science*, Dubai, UAE, 19-21 December 2011, pp. 37-48, Springer, 2011.

[2] **I. Vulić**, W. De Smet, and M.-F. Moens. "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," *Information Retrieval*, vol. 16, no. 3, pp. 331-368, Springer, 2013.

[3] M.-F. Moens and **I. Vulić**. "Monolingual and cross-lingual probabilistic topic models and their application in information retrieval," in *Proceedings of the 35th European Conference on Information Retrieval (ECIR)*, vol. 7814 of *Lecture Notes in Computer Science*, Moscow, Russian Federation, 24-27 March 2013, pp. 875-878, Springer, 2013.

# 12

# MuPTM and (Cross-Lingual) Relevance Modeling

*The problems are solved, not by giving new information, but by arranging what we have known since long.*

— Ludwig Wittgenstein

## 12.1 Introduction

In this chapter, we continue the exploration of the answers relevant to our research question RQ4. (i) How to represent multilingual content, that is, documents written in different languages in a structured and coherent way, regardless of their actual language? (ii) How to perform the effective retrieval of information (monolingually and across languages) that relies on such language-independent and language pair independent representations?

While the previous chapter has already introduced a probabilistic framework for constructing monolingual, cross-lingual and multilingual information retrieval models, in this chapter we extend the framework and explore the potential of multilingual probabilistic topic modeling within the probabilistic *relevance modeling* framework for both monolingual and cross-lingual IR.

In this chapter, we combine the answers to the raised questions (i) and (ii) into a powerful and robust language-pair independent unified framework for retrieval which is the extension of our probabilistic framework from the previous

chapter. In chapters 4 and 11, it has already been shown how to provide an answer to question (i), that is, how to represent documents as mixtures of latent cross-lingual topics. These high-level structured document representations by means of MuPTM are effectively language-agnostic. Moreover, all previous work that explored the potential of probabilistic topic models for information retrieval in both monolingual [323, 330] and cross-lingual settings (e.g., our work reported in the previous chapter) dealt with only simpler *query likelihood models*. In order to satisfy the requirements from question (ii), we opt for the more complex and robust probabilistic *relevance-based retrieval framework* [163, 162]. Additionally, as in the previous chapter, we aim to build new retrieval models which do not rely on parallel corpora, end-to-end SMT systems trained on parallel data, or hand-crafted external translation dictionaries, since such translation resources are unavailable or limited for numerous language pairs and domains. We make several important contributions along these lines:

(i) We present a novel way of estimating relevance models by means of a multilingual topic model in both monolingual and cross-lingual settings. The estimation is performed without any additional translation resource, while previous estimation techniques for cross-lingual relevance models critically relied on either a machine-readable bilingual dictionary or an in-domain parallel corpus [162], not available for many languages and domains.

(ii) We demonstrate that by our estimation procedure we create a unified formal framework that does not make any conceptual distinction between monolingual retrieval and CLIR. The proposed framework combines the strength and robustness of relevance modeling (e.g., its implicit query expansion and disambiguation) with the strength of MuPTM (e.g., shallow semantic analysis of documents, representation by means of language-independent latent cross-lingual concepts/topics).

The reported results from the experiments on the standard CLEF datasets show the validity of our unified approach as: (1) Relevance modeling clearly benefits from the additional knowledge coming from a probabilistic topic model, and it is visible in both monolingual and cross-lingual retrieval settings; (2) Cross-lingual relevance models estimated by means of a multilingual topic model produce results which are better than or comparable to several strong monolingual baselines; (3) Cross-lingual relevance models may be estimated by using only comparable user-generated data, which is especially important for language pairs and domains that lack readily available machine-readable bilingual dictionaries or parallel corpora.

The remainder of the chapter is structured as follows. First, we formally define a relevance model (sect. 12.2), provide a short overview of estimation techniques for the relevance model, and present our novel estimation technique (sect. 12.3).

Following that, we present our experimental setup in sect. 12.4, and evaluate our new MuPTM+RM-based retrieval models and show their validity in the monolingual and cross-lingual subject search from the CLEF data as in the previous chapter (sect. 12.5). Finally, we summarize our conclusions and provide directions for future work in sect. 12.6.

Again, throughout this chapter, the modeling process in the cross-lingual setting will be described. The modeling in the monolingual setting may be observed as an easier special case.

## 12.2 A Short Introduction to Relevance Modeling

Assume that we have two disjunct classes of documents according to their relation to a submitted query $Q$: $R_Q$ represents the class of documents relevant to the user's query $Q$, while $NR_Q$ represents the class of non-relevant documents [256, 257]. In a perfect scenario when the classes $R_Q$ and $NR_Q$ are already known for a submitted query, it is possible to rank documents from a target document collection according to their similarity to the "relevant class" $R_Q$. There has been a large body of research devoted to building IR models relying on the *explicit models of relevance* (e.g., [300, 258]). However, the main obstacle here is the challenge of estimating the relevant class which is difficult due to a lack of training data already labeled with classes according to their relevance assessments. Since we live in an imperfect world, in a typical retrieval setting we are given a query and a large target document collection without any indication of which documents might be relevant to the submitted query. In this chapter, we adopt the LM-inspired [53] approach and terminology from Lavrenko et al. [163, 162], and provide the following formal definition of a *relevance model*:

**Definition 12.1. Relevance model.** The term relevance model (RM) addresses a probability distribution that specifies the expectancy that any given word is observed in a set of documents relevant to the issued query. The relevance model comprises all probability scores $P(w|R_Q)$ which denote a probability of *sampling* a word $w$ from the documents relevant to the query $Q$.

The relevance modeling framework [163] assumes that both a query and its relevant documents are random samples from an underlying relevance model $R_Q$, where the sampling process could be different for queries and documents (see fig. 12.1).

If we operate in the cross-lingual setting, we need to estimate a *cross-lingual relevance model* (CRM) which comprises all probability scores $P(w^T|R_Q^T)$. In other words, we have to: (i) estimate a set of relevant documents in the target

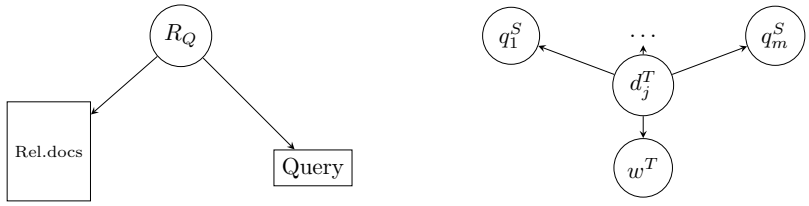



Figure 12.1: Queries and relevant documents are random samples from an underlying relevance model $R_Q$ (left). A dependence network for the estimation of the joint probability $P(w^T, q_1^S, \ldots, q_m^S)$ (right): The query words $q_1^S, \ldots, q_m^S$ and the words $w^T$ are sampled independently and identically from a distribution representing the document $d_j^T$.

language $R_Q^T$ given a query $Q^S$ in the source language, and (ii) provide a *sampling procedure* both for query terms given in the source language and words given in the target language (see again fig. 12.1). We will show that it is possible to estimate a cross-lingual relevance model by means of a multilingual probabilistic topic model.

## 12.3 Cross-Lingual Estimation of a Relevance Model

### 12.3.1 Prior Work

As already discussed in more detail in chapter 11, in recent years, numerous language modeling techniques were proposed to deal with the task of cross-lingual information retrieval. The common approach is to perform a word-by-word translation of a query in the source language to the target language by means of word translation probabilities [15, 127, 22, 319]. The translation probabilities are obtained from a bilingual dictionary or are induced from parallel corpora using alignment models for statistical machine translation [37, 229], or association measures based on hypothesis testing [215]. However, CLIR models relying on cross-lingual relevance models [162] proved to be superior compared to these models in CLIR tasks [162], but their estimation still critically relies on an external translation resource such as a bilingual dictionary or an in-domain parallel corpus. Additionally, monolingual IR models built within the relevance modeling framework also outperform other monolingual IR models [163, 330].

### 12.3.2 Approximating a Cross-Lingual Relevance Model

The ranking of documents $d_j^T \in \mathcal{DC}^T$, where $\mathcal{DC}^T$ again stands for the target document collection given in the target language, could be achieved if one had a way to estimate the cross-lingual relevance model of the source query $Q^S$. In other words, we need to estimate the set of documents $R_Q^T$ relevant to $Q^S$, and estimate the set of probabilities $P(w^T|R_Q^T)$ for each word $w^T \in V^T$. Relevance models serve as a powerful and robust retrieval framework due to its implicit massive query expansion (since the value $P(w^T|R_Q^T)$ is calculated for each $w^T \in V^T$, and the original query is replaced with a distribution over the entire target language vocabulary) and its implicit disambiguation.

Here, we face two major problems in the cross-lingual setting: (1) We typically do not possess any knowledge of which documents comprise the set $R_Q^T$ (see sect. 12.2), (2) We have to bridge the gap between different languages, and model the concept of sampling of a source language query term from a target language document. In order to estimate the relevance model in the absence of any prior knowledge about the set $R_Q^T$, we follow the heuristic presented by Lavrenko et al. [163, 162]:

$$P(w^T|R_Q^T) \approx P(w^T|Q^S) = \frac{P(w^T, q_1^S, \ldots, q_m^S)}{P(q_1^S, \ldots, q_m^S)} \tag{12.1}$$

The probability $P(w^T|Q^S)$ denotes the chance to observe a target language word $w^T$, with respect to a set of underlying distributions $\mathcal{U}$ from which the words are sampled, conditioned on observing $m$ source language words $q_1^S, \ldots, q_m^S$ that constitute the query $Q^S$ in the source language. The set $\mathcal{U}$ is typically the target document collection $\mathcal{DC}^T$ [162].

Further, Lavrenko and Croft [163] propose two methods for estimating the joint probability $P(w^S, q_1^S, \ldots, q_m^S)$ in the monolingual setting when $w^S, q_1^S, \ldots, q_m^S \in V^S$. These two methods differ in their independence assumptions. In this chapter, similar to [162], we opt for the simpler method and adapt it to the cross-lingual setting. This estimation principle assumes that the query terms $q_1^S, \ldots, q_m^S$ and the words $w^T$ are sampled identically and independently from a unigram distribution representing a document model of a document $d_j^T$ (see fig.

12.1). The estimate is then computed as follows:

$$P(q_1^S, \ldots, q_m^S) = \sum_{d_j^T \in \mathcal{DC}^T} P(d_j^T) \left( \prod_{r=1}^{m} P(q_r^S | d_j^T) \right) \tag{12.2}$$

$$P(w^T, q_1^S, \ldots, q_m^S) = \sum_{d_j^T \in \mathcal{DC}^T} P(d_j^T) \left( P(w^T | d_j^T) \prod_{r=1}^{m} P(q_r^S | d_j^T) \right) \tag{12.3}$$

### 12.3.3 Making the Model Tractable

The previous estimation model assumes that eq. (12.3) is calculated over every document $d_j^T \in \mathcal{DC}^T$, and it is repeated for each word $w^T \in V^T$. In case of a large vocabulary and a huge target document collection, the estimation is almost computationally infeasible. Therefore, we need an approximate, computationally tractable estimation of the probability $P(w^T | R_Q^T)$. The probability $P(w^T | R_Q^T)$ may be decomposed as:

$$P(w^T | R_Q^T) \approx P(w^T | q_1^S, \ldots, q_m^S) = \sum_{d_j^T \in \mathcal{DC}^T} P(w^T | d_j^T) P(d_j^T | q_1^S, \ldots, q_m^S) \tag{12.4}$$

The posterior probability $P(d_j^T | q_1^S, \ldots, q_m^S)$ then may be expressed as:

$$P(d_j^T | q_1^S, \ldots, q_m^S) = \frac{P(d_j^T) \prod_{r=1}^{m} P(q_r^S | d_j^T)}{\sum_{d_l^T \in \mathcal{DC}^T} P(d_l^T) \prod_{r=1}^{m} P(q_r^S | d_l^T)} \tag{12.5}$$

A relevance model under this estimation method is actually a linear mixture of distributions from $\mathcal{DC}^T$, where each distribution representing a document model $d_j^T$ is weighted by its posterior probability (see eq. (12.5)) of generating the query. This probability has negligible near-zero values for all but a few documents $d_j^T$ from the target document collection. These target documents are exactly the documents that obtain the highest scores for the source query $Q^S$. In order to speed up the retrieval process, we have decided to calculate eq. (12.4) over only the top $TOP$ target documents for the query $Q^S$ (e.g., initially ranking them by any query likelihood model as described in the previous chapter, e.g., by MuPTM-Basic, MuPTM-DM or MuPTM-SemLex), instead of calculating eq. (12.3) over the entire collection [162, 161].

Expressing the relevance model in terms of eq. (12.4) and eq. (12.5) displays another advantage [162]. Namely, one could relax the strict probabilistic interpretation of the posterior $P(d_j^T | q_1^S, \ldots, q_m^S)$ and use any heuristic estimate with non-negative values that sum up to 1. In other words, the relevance model

could be constructed from any initial ranked list of documents, which does not have to be built by a probabilistic initial model.

Finally, we still have to model the probabilities that constitute eq. (12.4) and eq. (12.5). $P(d_j^T)$ denotes some prior distribution over the dataset (i.e., the set of distributions) which is usually assumed as uniform. For estimation of the probabilities $P(w^T|d_j^T)$ and $P(q_r^S|d_j^T)$ which constitute the core of the cross-lingual relevance model, we will utilize the knowledge from a multilingual probabilistic topic model.

### 12.3.4 Estimating CRM by MuPTM

Again, assume that we have a multilingual topic model trained on a bilingual corpus. It is then possible to infer the model on the target document collection $\mathcal{DC}^T$, that is, each $d_j^T \in \mathcal{DC}^T$ may be represented by per-document topic distributions with scores $P(z_k|d_j^T)$. Additionally, each word $w$, regardless of its language, is assigned a probability $P(w|z_k)$. If words $q_r^S \in V^S$ and $w^T \in V^T$ were observed during the training of the multilingual topic model, they will receive the corresponding scores $P(q_r^S|z_k)$ and $P(w^T|z_k)$. We can now easily calculate the probabilities $P(w^T|d_j^T)$ and $P(q_r^S|d_j^T)$ using the shared latent cross-lingual topic space:

$$P(w^T|d_j^T) = \sum_{k=1}^{K} P(w^T|z_k)P(z_k|d_j^T) \quad P(q_r^S|d_j^T) = \sum_{k=1}^{K} P(q_r^S|z_k)P(z_k|d_j^T) \tag{12.6}$$

Again, recall that there is conceptually no difference between the monolingual calculation and the calculation across languages. In other words, under the assumptions made in this thesis, we have constructed a unified framework for both monolingual and cross-lingual relevance modeling. Additionally, note that the estimation of the probabilities $P(q_r^S|d_j^T)$ follows exactly the estimation already introduced with our MuPTM-Basic model from chapter 11 (see alg. 11.1 in sect. 11.3.1), while the estimation of the probabilities $P(w^T|d_j^T)$ occurs monolingually and therefore follows exactly the estimation with our MoPTM-Basic model (see sect. 11.5.5), which is in turn only a special degenerate case of the MuPTM-Basic model in the monolingual setting. Following this line of thinking, instead of MuPTM/MoPTM-Basic, we may use any other model from chapter 11 to estimate these probabilities.

### 12.3.5 Final Retrieval Model

We may now summarize the entire retrieval process which combines the knowledge from multilingual probabilistic topic models with the framework of relevance modeling and is able to operate both monolingually and cross-lingually. The retrieval process is summarized by alg. 12.1. Note that since

**Algorithm 12.1**: RELEVANCE MODELING WITH MUPTM

**Input** : bilingual training corpus $\mathcal{C} = \mathcal{C}_S \cup \mathcal{C}_T$, target document collection $\mathcal{DC}^T$, user query $Q^S$ ;

1: **train** the model on a (usually general-domain) training corpus and learn per-topic word distributions $\phi$ and $\psi$, and per-document topic distributions ;

2: **infer** the trained model on $\mathcal{DC}^T$ in the target language $L_T$ and obtain per-document topic distributions $\theta^T$ for all documents in $\mathcal{DC}^T$ ;

3: **perform** the *first retrieval round* with a query-likelihood MoPTM-based (monolingual setting) or MuPTM-based (cross-lingual setting) retrieval model (e.g., using MuPTM-Basic/DM/SemLex) or any other initial retrieval model;

4: **retain** only $TOP$ top scoring documents from the previous step as pseudo-relevant documents; estimate the probability scores $P(q_r^S|d_j^T)$ and $P(w^T|d_j^T)$ using the estimation procedure from sect. 12.3.4, but only over these $TOP$ documents (see sect. 12.3.3) ;

5: **estimate** the relevance model $P(w^T|R_Q^T)$ for each $w^T \in V^T$ by calculating eq. (12.4) and eq. (12.5) over these $TOP$ documents ;

6: **perform** the *second retrieval round* over the entire collection $\mathcal{DC}^T$ or, in a real-life retrieval setting, it is more common and less time-consuming to perform only the re-ranking of the top best scoring documents retrieved in the first retrieval round; each document $d_j^T$ is assigned a score that is the relative entropy between a relevance model $R_Q^T$ and a target document model $d_j^T$:

$$KL(R_Q^T||d_j^T) = \sum_{w^T \in V^T} P(w^T|R_Q^T) \log \frac{P(w^T|R_Q^T)}{P(w^T|d_j^T)} \tag{12.7}$$

7: **rank** documents in terms of their increasing relative entropy score ;

**Output**: $DRank(Q^S) \rightarrow$ the ranking of all documents from $\mathcal{DC}^T$ according to their relevance to $Q^S$ ;

documents have the same language-independent representation given by the distributions over cross-lingual topics, it allows for retrieving documents from a target collection given in multiple languages. In other words, documents relevant to the query may be in different languages, and the proposed framework is able to process it in a uniform way.

## 12.4 Experimental Setup

### 12.4.1 Training Data, Test Data and Evaluation Metrics

Training collections are exactly the same as in the previous chapter (see sect. 11.4.2). We have trained the BiLDA model with $K = 1000$ topics and symmetric hyper-parameters $\alpha = 50/K$, $\beta = 0.01$ [278] on the Wiki+EP bilingual corpus. In addition, we have trained the LDA model with exactly the same model parameters on the Dutch side of the bilingual corpus in order to also test our framework in the task of monolingual ad-hoc retrieval.

We again use the CLEF data for testing and evaluate our model in English-to-Dutch cross-lingual information retrieval and Dutch monolingual retrieval. Therefore, we use the NC+AD target document collection associated with EN 2001-2003 cross-lingual CLEF campaigns (which we now, for the sake of simplicity, name CLEF 2001-2003), remove stop words from queries and documents and retain only queries with at least one relevant document in the target collection (see tab. 11.2 and tab. 11.3 from sect. 11.4.3).

Evaluation metrics are again MAP and 11-pt recall-precision curves.

### 12.4.2 Models for Comparison

Our probabilistic IR framework designed in chapter 11 and this chapter allows for constructing a myriad of different retrieval models as it allows for choosing different approaches in each step of the retrieval process (e.g., one may use different models to estimate probabilities $P(w^T|R_Q^T)$ and $P(q_i^S|R_Q^T)$, one may use different models to perform the first retrieval round in alg. 12.1, one may use different dependency networks in relevance modeling). Here, we have opted for MoPTM-DM (monolingual setting) and MuPTM-DM (cross-lingual setting) in the first retrieval round. Since our main goal is to demonstrate that by combining MuPTM and the relevance modeling framework we are able to build more robust and more effective models of monolingual and cross-lingual ad-hoc retrieval, we have decided to compare these retrieval models:

(1) Monolingual relevance model estimated using only the document model representation (see eq. (11.3) in sect. 11.3.2). The model is estimated according to [163]. It was used before as a strong monolingual baseline [162, 330] (the MRM+DM model).

(2) Monolingual query likelihood MoPTM-based retrieval model that linearly combines the document model (DM) and the monolingual topic model (MoPTM)

representation as in eq. (11.6) in sect. 11.3.2 (MQL+MoPTM-DM). It is the best scoring model in [323].

(3) Monolingual relevance model estimated using both the DM and the MoPTM representation (again according to eq. (11.6)). Our goal is to test whether combining relevance modeling and topic modeling in the monolingual setting also leads to a better model and, consequently to a stronger monolingual baseline (MRM+MuPTM-DM).

(4) Cross-lingual query likelihood MuPTM-based retrieval model that linearly combines the DM and the multilingual probabilistic topic model (MuPTM) representation as given by eq. (11.6) in sect. 11.3.2 (CQL+MuPTM-DM).

(5) Cross-lingual translation model which uses *Google Translate* to perform an automatic translation of the original query as formulated by [327], and then effectively performs monolingual retrieval using both the DM and the monolingual topic model representation as in the previous MQL+MoPTM-DM model (CQL+GT+MoPTM-DM, see sect. 11.5.5).

(6) Cross-lingual relevance model estimated by eq. (12.4), eq. (12.5) and eq. (11.6), which combines both the DM and the MuPTM representation within the cross-lingual relevance modeling framework (CRM+MuPTM-DM).

### 12.4.3 Parameters

The parameter of the Dirichlet prior is again set to the standard value of $1,000$ (see sect. 11.4.4). As in the previous chapter, we fix the value $\lambda = 0.3$ for MuPTM-DM and MoPTM-DM. To estimate the relevance model of a query in all models, we use $TOP = 50$ best scoring documents from the first retrieval round, according to Lavrenko and Allan [161]. They present the full analysis of the impact of reducing the number of documents to only top $TOP$ documents considered for expansion on the speed and effectiveness of relevance-based retrieval models.

## 12.5 Experiments, Results and Discussion

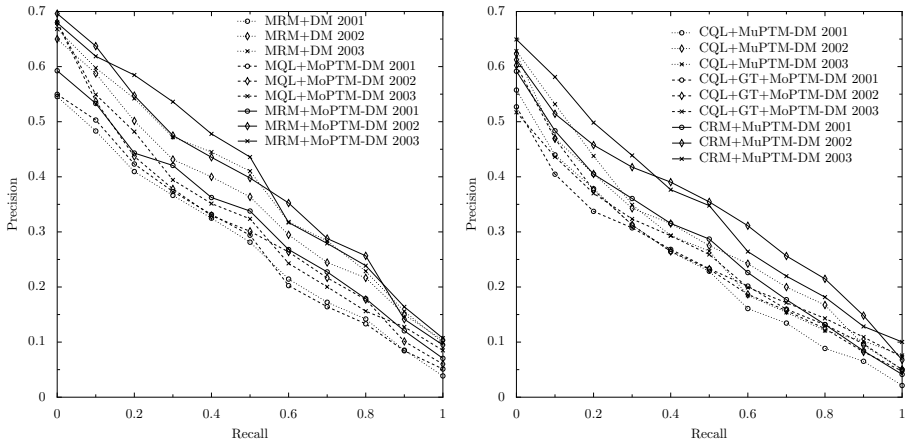
The MAP scores over both retrieval tasks are displayed in tab. 12.1. Additionally, 11-pt recall-precision curves are presented in fig. 12.2a and fig. 12.2b which respectively compare our monolingual and cross-lingual retrieval models. Based on these results, we may derive several interesting conclusions:

| Model | CLEF-2001 | | CLEF-2002 | | CLEF-2003 | |
|---|---|---|---|---|---|---|
| **MRM+DM (○)** | $0.2637^{\bullet\diamond\clubsuit}$ | | $0.3340^{\star\bullet\diamond\clubsuit}$ | | $0.3539^{\bullet\diamond\clubsuit}$ | |
| **MQL+MoPTM-DM (⋆)** | $0.2603^{\bullet\clubsuit}$ | -1% | $0.2891^{\circ\bullet\diamond\clubsuit\triangle}$ | -13% | $0.3262^{\bullet\clubsuit}$ | -8% |
| **MRM+MoPTM-DM (•)** | $0.3042^{\circ\star\diamond\clubsuit}$ | +15% | $0.3709^{\circ\star\diamond\clubsuit\triangle}$ | +11% | $0.3836^{\circ\star\diamond\clubsuit\triangle}$ | +8% |
| **CQL+MuPTM-DM (◇)** | $0.2275^{\circ\bullet\triangle}$ | -14% | $0.2683^{\circ\star\bullet\triangle}$ | -20% | $0.2783^{\circ\bullet\clubsuit\triangle}$ | -21% |
| **CQL+GT+MoPTM-DM (♣)** | $0.2296^{\circ\star\bullet\triangle}$ | -13% | $0.2401^{\circ\star\bullet\triangle}$ | -28% | $0.2443^{\circ\star\bullet\diamond\triangle}$ | -31% |
| **CRM+MuPTM-DM (△)** | $0.2689^{\diamond\clubsuit}$ | +2% | $0.3372^{\star\bullet\diamond\clubsuit}$ | +1% | $0.3351^{\bullet\diamond\triangle}$ | -5% |

Table 12.1: MAP scores on the CLEF monolingual and cross-lingual retrieval task with English (and Dutch) queries and Dutch document collection. All relative performances are given with respect to the baseline MRM+DM model performance. Each model is also assigned a unique symbol. The symbols indicate statistically significant differences between the MAP scores in each campaign of every two models to which these symbols are assigned. We use the two-tailed t-test ($p < 0.05$).

(i) The general important conclusion is that combining the advantages of probabilistic topic modeling and relevance modeling leads to a better performance of probabilistic language models for retrieval in both the monolingual and the cross-lingual settings. The MRM+MoPTM-DM model which uses both the original document representation and the monolingual probabilistic topic model representation outperforms a strong monolingual baseline (the MRM+DM model) which also relies on relevance modeling, but utilizes only the original document representation to estimate the relevance model. Therefore, the MRM+MoPTM-DM should be used as a stronger monolingual baseline.

(ii) Comparisons between MRM+MoPTM-DM and MQL+MoPTM-DM on one hand, and MRM+MoPTM-DM and MRM+DM on the other hand, reveal that both relevance modeling and probabilistic topic modeling are significant factors in constructing high quality retrieval models. The most powerful and robust retrieval models are built by combining the two. Another important remark is that all previous work on topic models in ad-hoc monolingual retrieval relied on in-domain corpora to train the models and learn the topical structure [323, 330] (i.e., they train on newswire corpora and perform retrieval on another newswire




(a) Monolingual retrieval (b) Cross-lingual retrieval

Figure 12.2: 11-pt recall-precision curves for all models over all campaigns. The positive synergy between probabilistic topic modeling and relevance modeling is clearly visible in both the monolingual setting and the cross-lingual setting. A similar relative performance is observed in the reverse retrieval direction (Dutch queries, English documents) and in the English monolingual retrieval task.

corpus). Here, we show that such models may also benefit from the topical knowledge coming from a general corpus such as Wikipedia.

(iii) In the cross-lingual setting, it is again visible that the CRM+MuPTM-DM model, which combines relevance modeling and two different representations of a document, outperforms the two other CLIR models by a significant margin. A cross-lingual probabilistic translation model (CQL+GT+MoPTM-DM) is not sufficient to fully capture the semantics of the query and disambiguate the query terms and knowledge is lost in the translation process. On the other hand, latent cross-lingual topics have an ability to capture the semantics of the query, as the query words are likely to be generated by particular cross-lingual topics and, consequently, a higher preference is assigned to documents dominated by these most likely topics in their topic representation.

(iv) Although latent cross-lingual topics serve as a bridge between two languages and as implicit query disambiguation tool, a simple query likelihood model such as CQL+MuPTM-DM [313] is still not sufficient to obtain results comparable to the monolingual retrieval models. However, by integrating the topical knowledge in the proposed cross-lingual relevance modeling framework, we are able to build a CLIR model (CRM+MuPTM-DM) that outscores the simple query likelihood CLIR model. The CRM+MuPTM-DM model is more complex and has a higher computational complexity, but it is more robust and effective.

(v) A comparison of the CRM+MuPTM-DM model with the monolingual baselines reveals that its performance is on a par with the MRM+DM model which does not rely on any topical knowledge, and it reaches up to 90% of the average performance of the MRM+MoPTM-DM model, which is conceptually the same model, but operating in the monolingual setting. We believe that CRM+MuPTM-DM displays an excellent overall performance, especially taking into account that it does not utilize any translation resource and relies only on a general non-parallel corpus for training.

## 12.6 Conclusions and Future Work

In this chapter, we have extended our work on theoretical (CL)IR models. We have proposed a unified framework for monolingual and cross-lingual information retrieval which combines the modeling advantages of multilingual probabilistic topic modeling and relevance modeling. While multilingual topic models have a capability to represent each document in a collection as a mixture of language-independent concepts, that is, cross-lingual topics, regardless of the actual language of the documents, relevance models additionally provide a robust framework for massive query expansion and disambiguation.

We have presented an estimation procedure for the relevance models by means of a multilingual topic model which relies only on general non-parallel data easily obtainable from the Web (e.g., Wikipedia articles), unlike all prior work that relied exclusively on in-domain parallel data or bilingual lexicons. Again, the proposed framework is generic, language pair independent and model-independent, as it allows for inputting any multilingual topic model that outputs the sets of per-topic word and per-document topic distributions in the relevance modeling framework. However, cross-lingual topics discovered in a general corpus might be too coarse-grained or non-relevant for certain documents in the target collection. In that case, the retrieval model is not completely able to capture the query semantics and transfer it across languages to perform effective retrieval. It is especially prominent with queries that have only one or two relevant documents. Therefore, one path in future work leads towards designing the estimation procedure that will solve the issue with such queries. Our goal in this chapter was to provide and describe the general theoretical aspects of the framework, which allows for building plenty other retrieval models sharing the same theoretical aspects.

We have conducted a thorough analysis of our models in the monolingual and cross-lingual ad-hoc retrieval tasks on the standard CLEF test collection. Our results show that the topical knowledge learned on a general corpus is useful when combined with the framework of relevance modeling in both monolingual and cross-lingual settings. Additionally, current state-of-the-art CLIR models that exploit the topical knowledge [323, 313] are outperformed by the models built within this novel framework.

## 12.7 Related Publications

[1] **I. Vulić** and M.-F. Moens. "A unified framework for monolingual and cross-lingual relevance modeling based on probabilistic topic models," in *Proceedings of the 35th European Conference on Information Retrieval (ECIR)*, vol. 7814 of *Lecture Notes in Computer Science*, Moscow, Russian Federation, 24-27 March 2013, pp. 98-109, Springer, 2013.

# 13

# Conclusions and Future Perspectives

*I am turned into a sort of machine for observing facts and grinding out conclusions.*

— Charles Darwin

We conclude by summarizing the presented work, restating its main research claims and contributions, and providing an outlook on future research.

## 13.1 Thesis Summary

Observing the enormous growth in global connectivity and the number of Web users worldwide, information becomes increasingly available in multiple different languages. Two important processes occur simultaneously - users generate more data in their native languages, while at the same time, they seek to access information which is not readily available in their native languages. We may justifiably say that today's world has truly become a data-driven multilingual environment. With respect to these observations, we have detected several major requirements around which the entire thesis work revolves. Here, we restate these requirements.

R1. Considering the large number of languages (and, consequently, language pairs), we need to represent multilingual text content, that is, words, phrases, documents written in different languages in a structured and coherent way, regardless of their actual language. In short, we require *language-*

*independent* and *language pair independent representations* of multilingual text content.

R2. The users should be able to understand the content which is not given in their native languages. A fundamental cross-lingual problem is the problem of *meaning* and cross-lingual *semantics*, which implies the problem of *translation*. We require widely applicable tools that are able to automatically detect a similar meaning of text units across a wide variety of language pairs and induce translational resources (e.g., bilingual lexicons) directly from the data itself.

R3. The users should be able to acquire knowledge and satisfy their information need from the relevant content which is not always given in their native languages. Therefore, we require tools which deal with another fundamental problem of *information retrieval* (monolingually and across languages). These tools should again be applicable to a wide variety of languages and language pairs.

R4. Besides the requirement of being *applicable to a wide spectrum of languages, idioms and domains*, the tools have to be *cheap*, *data-driven* and *effective*.

The complete dissertation is structured around these requirements, and the goal of this thesis has been to enlighten, address and tackle the major challenges arising from them (i.e., how to address the problems of content representation, meaning, translation and information search and retrieval across languages). In practice, (1) we demonstrate how to obtain language-independent and language pair independent representations of words and documents (part I of the thesis, related to requirements R1 and R4) (2) we propose a new statistical approach to extracting bilingual lexicons from parallel corpora (part II, related to requirement R2), (3) we introduce and describe a new fully corpus-based language pair independent framework for modeling semantic similarity and automatically inducing bilingual lexicons from comparable corpora (part III, related to requirements R2 and R4), (4) we introduce and describe a new language pair independent probabilistic framework for cross-lingual information retrieval (part IV, related to requirements R3 and R4).

In part I, we present a full systematic overview of multilingual probabilistic topic modeling (MuPTM). First, in chapter 3, a short introduction to multilingual text resources is given. We list and discuss two main types of multilingual text data: parallel and comparable corpora. We discuss their structure, properties and differences, and motivate the usage of both types in cross-lingual applications. In

chapter 4 we discuss the MuPTM framework in detail. The MuPTM framework is utilized to induce latent cross-lingual topics/concepts from raw multilingual text data, and is able to learn these latent topics even from non-parallel data. We discuss the main modeling premises behind the MuPTM framework, and show how to train, infer, evaluate and use multilingual topic models. This chapter also introduces the notion of per-topic word distributions and per-document topic distributions. These distributions are utilized to provide structured and uniform representations of words and documents regardless of their actual language.

Part II deals with extracting term translations, that is, inducing bilingual lexicons from parallel corpora. The proposed approach to bilingual lexicon extraction (BLE) described in chapter 5 is fully data-driven and uses only internal distributional evidence derived solely from the given parallel data. Our approach introduces the idea of sub-corpora sampling which relies on the paradigm of data reduction instead of data augmentation. We present a new BLE algorithm called SampLEX and show that this new algorithm outperforms all other state-of-the-art BLE models from parallel data. The power of the algorithm lies in the idea of data removal, and the algorithm exhibits a clear advantage over all other BLE models when dealing with low-frequency words, indirect associations and parallel data of limited size. We have also employed the algorithm in one of the modules of the full computer-assisted translation (CAT) tool. The ultimate goal of the so-called TermWise tool is to facilitate translators the complete translation process dealing with jargon heavy domain-specific terminology. In specific, the algorithm is used to produce translation candidates for domain-specific terms in the Belgian legislative domain. The CAT tool and the role of the SampLEX algorithm in the tool are presented in a nutshell in appendix A.

Part III discusses the models of cross-lingual semantic similarity (ambitiously tackling the problem of meaning and its preservation across languages) and, consequently, bilingual extraction in a more difficult setting where one has only comparable corpora in possession. Part III introduces a new framework for modeling cross-lingual semantic similarity which relies on the MuPTM-based representations of words and text units beyond word level (from part I). We make several important contributions to the field of distributional semantics.

First, in chapter 6 we review models of cross-lingual semantic similarity and then motivate the work in this cheap minimalist setting where one possesses only comparable data to model semantic similarity. The first set of new MuPTM-based models of cross-lingual similarity is presented in the chapter, and we also introduce and test the utility of topic space pruning. Our results in the task of bilingual lexicon extraction reveal several important phenomena. We observe that MuPTM-based models of similarity outperform the variant models which use standard monolingual LDA trained on concatenated document pairs. We

also observe the utility of translating the problem of similarity from the original word-document space to the lower-dimensional word-topic space as our models of similarity are both faster and more effective. We also report on the utility of topic space pruning, that is, reducing high-dimensional MuPTM-based word representations to only a selection of most important semantic axes leads to additional improvements in both final scores in the task of bilingual lexicon extraction, and execution time.

In chapter 7 we present an extension of the basic modeling framework targeted towards selecting only highly reliable translation pairs. The selection of only confident translation pairs is especially important in a setting dealing with noisy comparable data. We have introduced a new precision-oriented algorithm called SelRel which is able to learn a subset of the confident translation pairs and disregard noisy translation pairs.

Chapter 8 deals with another extension of our statistical framework for modeling cross-lingual semantic similarity. We introduce an idea of computing semantic similarity of words as the similarity of their semantic responses. We define semantic responding and free word association and draw direct analogies to research in cognitive science. We translate the problem of semantic similarity from a cross-lingual semantic space spanned by latent cross-lingual topics to a semantic space spanned by all vocabulary words in both languages which act as semantic responses. Our results in the BLE task show that this new semantic space yields more robust and more effective models of semantic similarity.

Chapter 9 presents a new bootstrapping framework for building a shared cross-lingual semantic space. In this chapter we have dissected the complete bootstrapping pipeline and have described all its key components: (1) the starting point or seed dimensions of the space, (2) the updating step with the confidence estimation and selection of new dimensions of the space, (3) convergence. We have made contributions in all steps of the pipeline. We have discussed how to use highly-confident translation pairs obtained initially by an MuPTM-based model of cross-lingual semantic similarity as seed dimensions of the space and have shown that the choice of seed dimensions matters for the bootstrapping process. Furthermore, we have introduced and tested several confidence estimation functions. Since the work introduced in this chapter is still preliminary and unravels new exciting ideas and research questions, we believe that further improvements of the bootstrapping framework are possible in future work.

Finally, in chapter 10, another extension of the framework is proposed. We propose new models of cross-lingual semantic similarity in context and demonstrate their utility in the task of word translation in context. We motivate the context-sensitive modeling of cross-lingual semantic similarity and show

that modeling the change of meaning is simply modulating word representations by means of MuPTM-based multinomial distributions. Our results reveal the utility of "contextualized" word representations and context-sensitive models of similarity. The work on context-sensitive models of cross-lingual similarity has not been investigated in the relevant literature and this chapter also provides plenty of directions for future work. It is possible to investigate other models of context aggregation and selection (e.g., including syntactic information or including neighboring words for larger context) and the influence of context scope, or build a similar framework that is able to operate with explicitly defined concept categories (as opposed to latent concepts), etc.

Part IV proposes a new language pair independent framework for cross-lingual information retrieval (CLIR). The framework is again supported by the knowledge of latent cross-lingual topics, but we also demonstrate how to include the evidence of semantically similar words cross-lingually (from part III) in new CLIR models. In chapter 11, we present the first set of new CLIR models starting from the basic MuPTM-based retrieval model which uses only document representations by means of per-document topic distributions. We thoroughly evaluate all retrieval models and report our findings. We observe that the sole MuPTM-based document representation is not sufficient to provide quality retrieval results, but combining different representations and different evidences in the retrieval process leads to more effective retrieval models. More robust CLIR models have been proposed in chapter 12. We have combined the MuPTM-based text representations within the relevance modeling framework for IR. The obtained results clearly show that this combination leads to more effective and more robust models of retrieval both monolingually and cross-lingually. In summary, due to its theoretical soundness and modularity, the proposed CLIR framework allows further extensions. It allows for building more models that will exploit different retrieval evidences in the retrieval process (e.g., external bilingual lexicons, cognates, combining representations obtained by different multilingual topic models).

## 13.2 Contributions

We have detected four major contributions of this thesis which have been presented in their respective parts of this thesis. We have to stress that the contributions span and unite a range of different research fields (e.g., data mining, computational linguistics, natural language processing, information retrieval), which makes it a true multidisciplinary thesis.

In part I, the first full systematic and comprehensible overview of a new multilingual text mining framework called multilingual probabilistic topic modeling has been provided. We present its theoretical background, modeling

assumptions, methodology and evaluation described all the way up from the conceptual and modeling level down to the mathematical level. Text representations by means of multilingual topic models have found numerous applications in various cross-lingual tasks. In this thesis we have addressed and investigated two fundamental applications: (i) cross-lingual semantic similarity (in part III) and (ii) cross-lingual information retrieval (in part IV).

In part II, a new approach to designing statistical models for bilingual lexicon extraction from parallel data has been introduced. The approach relies on the paradigm of sampling, data reduction and utilizing low-frequency events/words. A new algorithm developed under these modeling premises is tailored towards selecting only confident translation pairs, and it shows its utility when dealing with low-frequency words and parallel data of limited size. We employ the same modeling principle to extract domain-specific term translations and build a full-fledged CAT tool where this algorithm is run in the background as one of the modules.

In part III, a new framework for modeling cross-lingual semantic similarity which relies on the notion of latent cross-lingual concepts/topics has been introduced, described and evaluated. This framework has been pioneered within the work conducted in this thesis, and the research related to this framework has unfolded in a series of contributions to the research field of distributional semantics. Both part II and part III make significant contributions to the field of natural language processing and its sub-field of multilingual NLP, with a special focus on distributional semantics and cross-lingual models of meaning.

In part IV, we have made a significant contribution to the field of information retrieval and its sub-field of cross-lingual information retrieval. We have proposed a novel approach to cross-lingual information retrieval and the construction of a new CLIR MuPTM-based framework for retrieval which allows for embedding many additional evidences in building novel retrieval models without any additional external resources such as machine-readable translation dictionaries or parallel corpora. The new models constructed within this framework are cheap and effective, and their unsupervised nature should make them effectively applicable to a variety of language pairs.

## 13.3 Future Work

The work discussed in this thesis has tackled some fundamental problems in multilingual NLP and IR (e.g., the problem of meaning, translation, and information retrieval). As a consequence, it has also opened up new research questions and perspectives. From a theoretical perspective, the thesis raises the question whether the concept of meaning may be transferred and shared

across-languages (e.g., by modeling cross-lingual semantic similarity in part III or by modeling sampling of a source language query from a target language document in part IV). From a more practical perspective, this thesis has shown that effective tools for cross-lingual text processing tasks may be built even from noisy user-generated multilingual content. This finding should also encourage similar approaches in the *multimodal setting*, where even a larger semantic gap between different modalities exists. Porting the cross-lingual tools from the multilingual setting to the multimodal setting and building cross-modal tools is a natural step. For instance, there is a pressing need to build cross-modal retrieval models (similar to our cross-lingual models) which will enable the search for information given in multiple different modalities (e.g., search for text, images, video or audio material). Some preliminary efforts which rely on the topic modeling concept have already been ignited in this research domain (e.g., [7, 87, 259, 38]).

Additionally, besides being a multilingual environment, the Web and the world of information are also locales for multiple idioms of the same language. For instance, the "language" of the social media consumers or typical end-users differs from the language of Wikipedia entries, online shops or legal terminology. Different domains also display different usage of language. Therefore, one line of future research also lies in studying and applying the models initially tailored for the multilingual setting within this *multi-idiomatic* setting. We have already made the first step in that direction as we have proven the utility of MuPTM-based text representations and the (CL)IR framework developed in this thesis in a new task of linking Pinterest users to relevant online shops based on the content the users post on their personal pages [342, 341]. Moreover, we have recently proposed a new *multi-idiomatic* topic model [316]. In the long run, the ultimate goal is to build effective tools which should be able to cross domains, modalities, languages and the idiomatic usage of the same language.

Finally, in the pure multilingual setting there are still multiple open directions for future research. One direction tackles the development of new multilingual topic models. As already discussed in sect. 4.7 in chapter 4, in the same fashion as with the natural and straightforward "LDA to BiLDA" extension, existing monolingual probabilistic topic models could be transferred into the multilingual setting. These extensions comprise, among others, the use of sentence information and word ordering to yield more coherent topic distributions over documents. The use of hierarchical topics (general super-topics connected with more focused sub-topics) in the multilingual setting is also worth investigating. Moreover, there is a need to develop multilingual probabilistic topic models that fit data which is less comparable and more divergent and unstructured than Wikipedia or news stories, where only a subset of latent cross-lingual topics overlaps across documents written in different languages. Additionally, the

more data-driven topic models should be able to learn the optimal number of topics dynamically according to the properties of training data itself, and clearly distinguish between shared and non-shared topics in a multilingual corpus.

A related line of future work addresses the applications of multilingual topic models in other more-specific cross-lingual tasks beyond the two fundamental tasks discussed in this thesis. Initial studies have shown that text representations by means of MuPTM are useful in document classification, keyword recommendation, systems for news clustering and summarization, transliteration, building multilingual data resources, etc.

Another related line of future research in the multilingual setting deals with utilizing other multilingual topic models in the frameworks discussed in this thesis. In this thesis, we have employed BiLDA as a basic multilingual model (similar to LDA in the monolingual setting) in all applications. However, the modularity of the proposed frameworks allows for "plugging in" any other multilingual topic model which outputs the two basic sets of distributions: per-document topic distributions and per-topic word distributions. Additionally, by designing more advanced topic models which are able to capture finer-grained redundancies in data, the proposed frameworks for modeling cross-lingual semantic similarity and information retrieval may exploit this latent knowledge and build finer-grained models of cross-lingual similarity and retrieval.

Additionally, some chapters have also opened new more specific research perspectives. For instance, in chapter 10 dealing with context-sensitive models of cross-lingual similarity we have touched upon the problem of semantic compositionality. One line of future research should strive towards modeling semantic similarity at different levels of text granularity. In other words, following the recent work on sentence similarity in the monolingual setting [3], it would be interesting to measure similarity of phrases, sentences and other text chunks cross-lingually. In chapter 9 we also present a new bootstrapping framework for modeling cross-lingual semantic similarity which may spark more research interest in such models, even beyond the specific tasks of cross-lingual semantic similarity and bilingual lexicon extraction (e.g., weakly supervised models in information extraction).

The models for bilingual lexicon extraction and information retrieval developed in this thesis and their output may also be tested as modules or sources of knowledge in larger, user-oriented NLP and IR systems (e.g., one such application is the CAT tool introduced in appendix A). Other possibilities include systems for statistical machine translation, question answering, search and summarization of large text collections, etc.

# A

# Appendix - TermWise CAT Tool

*There is no real ending. It's just the place where you stop the story.*

— Frank Herbert

In this appendix, we present *TermWise*, the final deliverable of the *TermWise* project (IOF-KP/09/001) in a nutshell. TermWise is a *computer-assisted translation* (CAT) tool that offers additional terminological support for domain-specific translations. Compared to existing CAT-tools, TermWise has an extra database, a *Term&Phrase Memory*, that provides context-sensitive suggestions of translations for individual terms and domain-specific expressions. The Term&Phrase Memory has been compiled by applying newly developed statistical knowledge acquisition algorithms to large parallel corpora, and one of the modules of the tool is the *N-gram matching module* (see fig. A.1) which relies on our SampLEX algorithm for bilingual lexicon extraction (see part II - chapter 5). Although the entire TermWise CAT tool is designed as language pair- and domain-independent, the tool was developed in a project with translators from the Belgian Federal Justice Department (FOD Justitie/SPF Justice) as end-user partners. Therefore the tool is demonstrated and evaluated in a case study of bidirectional Dutch-French translation in the legal domain. In this appendix, we provide a short description of the CAT tool and its core components, but we do not provide all details of the entire framework. The main goal is to demonstrate the utility of our SampLEX algorithm in such a practical application.

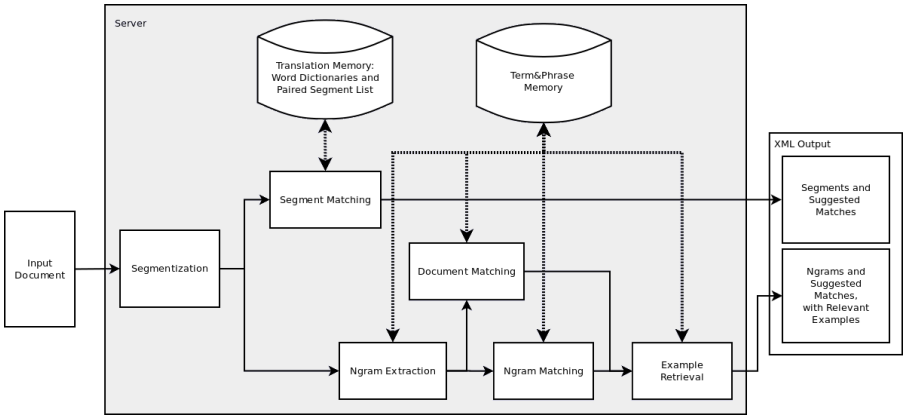



Figure A.1: TermWise CAT tool: an overview of its architecture.

## A.1 Why Term&Phrase Memory in a CAT Tool?

Like other domain-specific translators, the translators at the Belgian Ministry of Justice are confronted with source texts full of domain-specific terminology which requires exact (as opposed to interpretative) translation and which even skilled translators need to check against a reference source once in a while. However, in the commercial CAT-tool used by the Ministry, the support for terminological look-up is quite limited. As with most CAT-tools, it does come with a Term Base functionality, but this type of terminological dictionary is initially empty and entries have to be added manually. Even a large organization like the Ministry cannot afford to invest much time in Term Base compilation. They acquired an externally compiled Term Base, but its coverage is limited and it contains no informative examples of the idiomatic usage of terms in contexts. Such proper phraseological usage of terms is especially important in legal language, where validity of a text depends on the usage of the appropriate formulae. Although the commercial tool's Translation Memory (TM) can automatically retrieve translation suggestions, its operating level of entire sentences or even paragraphs is too course-grained for finding examples of individual words and phrases. A concordancer does allow for a manual look-up a specific expression, but occurrences are not sorted for relevance, nor do they come with metadata about the source document that could allow translators to assess its relevance and reliability. Additionally, the TM only keeps track of the Ministry's in-house translations, and does not include the vast body of relevant bilingual legal documents translated at other departments. The translators therefore often end up doing Google searches for terms and phrases in open online legal document repositories to check previous translations in specific contexts. However, also

here, the relevance of the search hits must be assessed manually. Based on this situation, we identified the following user needs: (1) Access to previous translations of domain-specific single and multi-word expressions; (2) Examples of usage in context to infer correct phraseology; (3) Information about the source documents of the translation examples; (4) Examples from all relevant documents that are available online; (5) Sorting the examples by relevance to the current assignment; (6) Easy access to the examples from within the tool.

To our knowledge, this combination of functionalities is not implemented in any existing CAT tool [249]. In TermWise they are grouped in a separate module, which we will call a *Term&Phrase Memory* (TPM), so that in principle this module can be integrated in existing CAT tools. The Term&Phrase Memory is basically an additional database accessible from within a CAT-tool's user-interface, next to the Translation Memory and Term Base. It contains domain-specific multi- and single word expressions and examples of their translations that were extracted automatically (using SampLEX) from a very large parallel corpus of online available, legal, bilingual documents.

## A.2 Knowledge Acquisition

In our legal case study, the relevant body of previous translations was defined as the laws, decrees, and official communications published in both French and Dutch by the Belgian state in the online version of Moniteur Belge/Belgisch Staatsblad (see sect. 3.2.1 in chapter 3) [307]. We also retrieved the source department (e.g. ministry, agency) for all documents. Both the Dutch and French corpus were POS-tagged with TreeTagger [267]. To extract domain-specific expressions and their translations, we followed the *extract-then-align* paradigm that is predominant in the literature on bilingual terminology extraction (e.g., see [60, 100, 69, 116, 181]). In this paradigm, domain-specific terms are first extracted for the two languages separately (domain-specific $N$-gram extraction) and then in a second step we seek for their candidate translations cross-lingually (cross-lingual $N$-gram matching). Although both tasks are well known in NLP and have many existing implementations, most current tools are geared towards delivering intermediate results for a Machine Translation system or further manual lexicon compilation. In the Term&Phrase Memory, however, the output has to be usable directly by end-users. We therefore developed our own statistical algorithms for term extraction and term alignment to accommodate the specific user needs above. The knowledge acquisition proceeded in two steps.

**Step I: Domain-Specific N-gram Extraction.** Following [150], we consider expressions of variable length as relevant for the legal domain. These do not only include single and multi-word terms that refer to legal concepts (typically noun phrases (NP)), but also phraseologies (e.g. typical verb-NP combinations), and formulaic expressions that can comprise entire clauses. The term extraction algorithm therefore considers $N$-grams of variable length without imposing predefined language-specific POS patterns as is the case in most term extraction algorithms. Instead, the relevancy of an $N$-gram is assessed based on its external *independence* and its internal *coherence*. Independence is the extent to which an $N$-gram can occur in different contexts. Coherence is the extent to which the lexemes within an $N$-gram tend to co-occur in an informational unit. Based on these properties, the extraction of relevant terms proceeds in three steps:

(1) In a preliminary step, the frequency of all sequential $N$-grams up to length 8 is calculated starting from adjectives, verbs and nouns as seeds.

(2) The independence of the $N$-grams is quantified by how frequent an $N$-gram is relative to each minimal expansion with one word. Following [56], $N$-grams are retained if they form local maxima, that is, they maximize frequency differences relative to the $N-1$ and $N+1$ grams in an $N$-gram expansion progression.

(3) The independent $N$-grams' internal coherence is measured by calculating the Mutual Information (MI) between elements of the $N$-gram. As a shortcut, only the MI between the first and last element of an $N$-gram is measured. Because the range of these lexical MI-values tends to be quite language specific, the MI values are calculated on the more schematic level of POS patterns. This results in cross-lingually comparable MI values and a single cut-off can be used as selection criterion in French and Dutch. In this case the cut-off was set at 6.7.

The algorithm is described in more detail in [65]. This extraction resulted in a *term list* of $649,602$ $N$-grams for French and $639,865$ $N$-grams for Dutch.

**Step II: Cross-Lingual N-gram Matching.** The goal of this step is to provide for each Dutch $N$-gram from the Dutch term list extracted in step I, a subset of likely candidate translations from the French $N$-grams term list (again extracted in step II) and vice versa. To find these term translations, we employ our SampLEX algorithm from chapter 5, but we adapt it to handle $N$-grams of variable length. In a pre-processing step, the aligned sentences in the corpus are represented as a *bag-of-terms* taken from the French and Dutch input term lists. Running SampLEX results for each Dutch $N$-gram/term in a list of French $N$-grams sorted by their respective translation probabilities, and vice versa. Also, we adapt the SampLEX algorithm so that it returns the document and sentence identifiers of each occurrence of a potential translation

pair in the corpus. As a post-processing step, a hard cut-off of the output lists of translation candidates is performed. Example output is displayed in tab. A.1. SampLEX focuses on the extraction of only highly-reliable translation pairs (see chapter 5) and that property is extremely important in a setting where a term extraction module (the module from step I) tends to "overgenerate" candidate terms.

| | |
|---|---|
| sur la proposition du conseil d' administration | |
| op voorstel van de raad van bestuur | Prob: 0.621 |
| op voordracht van de raad van bestuur | Prob: 0.379 |
| 16 mai 1989 et 11 juillet 1991 | |
| 16 mei 1989 en 11 juli 1991 | Prob: 1.0 |
| sur la proposition du ministre | |
| de voordracht van de minister | Prob: 0.481 |
| op voorstel van de minister | Prob: 0.111 |
| op voordracht van de minister | Prob: 0.074 |
| ... | ... |

Table A.1: Example output of the SampLEX algorithm operating with $N$-gram candidate terms. Translation direction: French to Dutch.

## A.3 Context-Sensitive Database Querying

Fig. A.1 shows the architecture of the TermWise tool. The system consists of a server, which handles translation requests, and a client, which issues the requests and displays the returned results in a graphical user interface. When handling a translation request, the server takes as input a text document and returns an XML file containing the segmentized document, translation suggestions for each segment, the $N$-grams found in the document, and translation suggestions for each $N$-gram together with context-sensitive annotated usage examples. The translation suggestions for segments correspond to the fuzzy matches from Term Memories in traditional CAT-tools and will not be further discussed here. Instead we will focus on handling of $N$-grams for the Term&Phrase functionality.

The Term&Phrase Memory consists of (a) a list of paired, sentence-aligned documents from the Belgian Official Journal annotated with their source department, and (b) a dictionary of the $N$-grams found in those documents. In the latter, each $N$-gram is associated with a list of translation candidates of a given translation probability, and each $N$-gram translation pair is associated with the list of documents and line numbers in which that translation is found.

When the server receives a translation request, the input document is first segmentized using the Alpino tokenizer [32]. $N$-grams are extracted from the segmentized input document by consulting the $N$-gram dictionary of the same language. A ranked list of similar corpus documents and their respective source departments is retrieved by calculating the number of $N$-grams in common with the input document.

$N$-gram translations to be suggested are chosen on the basis of the given translation probabilities and on document similarity. The list of documents that are similar to the input document is compared with the list of documents for each $N$-gram translation pair. The relevance value for an $N$-gram translation pair is determined by a weighted interpolation of its given translation probability and the cosine similarity of the highest-ranking document on its list (based on a "set of $N$-grams" vector space model). If the relevance value exceeds a configurable threshold, that $N$-gram translation pair is displayed and suggested to the user. Example sentences are extracted from the highest-ranking document and from other high-ranking documents from the same source department.

## A.4 Evaluation

The TermWise tool and its utility has been qualitatively evaluated by two end-user groups. In November 2013, students of legal translation at the KU Leuven, campus Antwerp became acquainted with the tool and reported their experiences. In January 2014, professional translators at the Belgian Ministry of Justice also assessed the usability of the tool. The qualitative evaluation makes use of observational data and a survey. First, legal translators are invited to make use of the tool to translate an unseen legal text and give comments and feed-back on the Term&Phrase Memory functionality as they are translating. Afterwards, they are also asked to fill in a survey on the general usability of the tool and the new functionality it offers. Results of this qualitative evaluation will be used to improve the tool's user-friendliness and to fine-tune the parameters of the knowledge acquisition algorithms and the context-sensitive search function.

## A.5 Related Publications

[1] K. Heylen, S. Bond, **I. Vulić**, D. De Hertog, and H. Kockaert. "TermWise: A CAT tool with context-sensitive terminological support, " accepted for publication in the *9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 26-31 May 2014, pp. 4018-4022, ELRA, 2014.

# Bibliography


[1] M. Adriani and C. J. van Rijsbergen, "Term similarity-based query expansion for cross-language information retrieval," in *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 1999, pp. 311–322.

[2] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in *Proceedings of the 10th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2009, pp. 19–27.

[3] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "*SEM 2013 shared task: Semantic textual similarity," in *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics (*SEM)*, 2013, pp. 32–43.

[4] A. Agresti, *Categorical Data Analysis, 2nd Edition.* Wiley, 2002.

[5] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, 2013, pp. 13–22.

[6] D. Andrade, T. Nasukawa, and J. Tsujii, "Robust measurement and comparison of context similarity for finding translation pairs," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 19–27.

[7] M. Andrews, G. Vigliocco, and D. Vinson, "Integrating experiential and distributional data to learn semantic representations." *Psychological Review*, vol. 116, no. 3, pp. 463–498, 2009.

[8] M. Apidianaki, "Unsupervised cross-lingual lexical substitution," in *Proceedings of the 1st Workshop on Unsupervised Learning in NLP*, 2011, pp. 13–23.

[9] M. Apidianaki, "LIMSI : Cross-lingual word sense disambiguation using translation sense clustering," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SEMEVAL)*, 2013, pp. 178–182.

[10] J. Arenas-García, A. Meng, K. B. Petersen, T. L. Schiøler, L. K. Hansen, and J. Larsen, "Unveiling music structure via PLSA similarity fusion," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2007, pp. 419–424.

[11] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009, pp. 27–34.

[12] W. Aziz and L. Specia, "Combining dictionaries and contextual information for cross-lingual lexical substitution," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 117–122.

[13] L. Azzopardi, M. de Rijke, and K. Balog, "Building simulated queries for known-item topics: An analysis using six European languages," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2007, pp. 455–462.

[14] L. Azzopardi, M. Girolami, and C. J. van Rijsbergen, "Investigating the relationship between language model perplexity and IR precision-recall measures," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2003, pp. 369–370.

[15] L. Ballesteros and W. B. Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1997, pp. 84–91.

[16] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: Are more languages better?" in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 28–36.

[17] M. Baroni and A. Lenci, "Distributional memory: A general framework for corpus-based semantics," *Computational Linguistics*, vol. 36, no. 4, pp. 673–721, 2010.

[18] M. Baroni and R. Zamparelli, "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010, pp. 1183–1193.

[19] P. Basile and G. Semeraro, "UBA: Using automatic translation and Wikipedia for cross-lingual lexical substitution," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 242–247.

[20] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[21] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal, "Bridging the lexical chasm: Statistical approaches to answer-finding," in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2000, pp. 192–199.

[22] A. Berger and J. Lafferty, "Information retrieval as statistical translation," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999, pp. 222–229.

[23] A. Bhattacharyya, "On a measure of divergence two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 199–209, 1943.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[25] W. Blacoe and M. Lapata, "A comparison of vector-based representations for semantic composition," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 546–556.

[26] D. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," in *Proceedings of the 16th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2003.

[27] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[28] D. M. Blei, K. Franks, M. I. Jordan, and I. S. Mian, "Statistical modeling of biomedical corpora: Mining the caenorhabditis genetic center bibliography for genes related to life span," *BMC Bioinformatics*, vol. 7, p. 250, 2006.

[29] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2003, pp. 127–134.

[30] D. M. Blei and J. D. McAuliffe, "Supervised topic models," in *Proceedings of the 21st Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 121–128.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[32] G. Bouma, G. Van Noord, and R. Malouf, "Alpino: Wide-coverage computational analysis of Dutch," *Language and Computers*, vol. 37, no. 1, pp. 45–59, 2001.

[33] J. Boyd-Graber and D. M. Blei, "Multilingual topic models for unaligned text," in *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009, pp. 75–82.

[34] J. L. Boyd-Graber and D. Blei, "Syntactic topic models," in *Proceedings of the 22nd Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 185–192.

[35] J. L. Boyd-Graber, D. M. Blei, and X. Zhu, "A topic model for word sense disambiguation," in *Proceedings of the Joint 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 1024–1033.

[36] T. Brants, "TnT: A statistical part-of-speech tagger," in *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP)*, 2000, pp. 224–231.

[37] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[38] E. Bruni, N.-K. Tran, and M. Baroni, "Multimodal distributional semantics," *Journal of Artificial Intelligence Research*, vol. 49, pp. 1–47, 2014.

[39] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[40] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior Research Methods*, vol. 39, no. 3, pp. 510–526, 2007.

[41] J. G. Carbonell, J. G. Yang, R. E. Frederking, R. D. Brown, Y. Geng, D. Lee, Y. Frederking, R. E, R. D. Geng, and Y. Yang, "Translingual information retrieval: A comparative evaluation," in *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*, 1997, pp. 708–714.

[42] M. Carpuat, "NRC: A machine translation approach to cross-lingual word sense disambiguation (SemEval-2013 task 10)," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SEMEVAL)*, 2013, pp. 188–192.

[43] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307, 2007.

[44] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Proceedings of the 23rd Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 288–296.

[45] D. Chen, Y. Xiong, J. Yan, G.-R. Xue, G. Wang, and Z. Chen, "Knowledge transfer for cross domain learning to rank," *Information Retrieval*, vol. 13, no. 3, pp. 236–253, 2010.

[46] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[47] P. A. Chew, B. W. Bader, T. G. Kolda, and A. Abdelali, "Cross-language information retrieval using PARAFAC2," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2007, pp. 143–152.

[48] K. W. Church and R. L. Mercer, "Introduction to the special issue on computational linguistics using large corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 1–24, 1993.

[49] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab, "Explicit versus latent concept models for cross-language information retrieval," in *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI)*, 2009, pp. 1513–1518.

[50] D. Clarke, "A context-theoretic framework for compositionality in distributional semantics," *Computational Linguistics*, vol. 38, no. 1, pp. 41–71, 2012.

[51] M. Collins, "Three generative, lexicalised models for statistical parsing," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL-EACL)*, 1997, pp. 16–23.

[52] M. Collins, "Head-driven statistical models for natural language parsing," *Computational Linguistics*, vol. 29, no. 4, pp. 589–637, 2003.

[53] W. B. Croft and J. Lafferty, *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, 2003.

[54] J. R. Curran, T. Murphy, and B. Scholz, "Minimising semantic drift with mutual exclusion bootstrapping," in *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 2007, pp. 172–180.

[55] I. Czarnowski and P. Jedrzejowicz, "Data reduction algorithm for machine learning and data mining," in *Proceedings of the 21st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE)*, 2008, pp. 276–285.

[56] J. F. da Silva, G. Dias, S. Guilloré, and J. G. P. Lopes, "Using LocalMaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units," in *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence (EPIA)*, 1999, pp. 113–132.

[57] I. Dagan, K. W. Church, and W. A. Gale, "Robust bilingual word alignment for machine aided translation," in *Proceedings of the Workshop on Very Large Corpora*, 1993, pp. 1–8.

[58] I. Dagan, L. Lee, and F. C. N. Pereira, "Similarity-based methods for word sense disambiguation," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1997, pp. 56–63.

[59] I. Dagan, F. C. N. Pereira, and L. Lee, "Similarity-based estimation of word cooccurrence probabilities," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1994, pp. 272–278.

[60] B. Daille, É. Gaussier, and J.-M. Langé, "Towards automatic extraction of monolingual and bilingual terminology," in *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, 1994, pp. 515–524.

[61] A. Darwiche, *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

[62] A. Darwiche, "Bayesian networks," *Communications of the ACM*, vol. 53, no. 12, pp. 80–90, 2010.

[63] D. Das and S. Petrov, "Unsupervised part-of-speech tagging with bilingual graph-based projections," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011, pp. 600–609.

[64] H. Daumé III and J. Jagarlamudi, "Domain adaptation for machine translation by mining unseen words," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011, pp. 407–412.

[65] D. De Hertog, "TermWise Xtract: Automatic term extraction applied to the legal domain," PhD, KU Leuven, 2014.

[66] W. De Smet, "Probabilistic graphical models for content representation and applications in monolingual and multilingual settings," PhD, KU Leuven, 2011.

[67] W. De Smet and M.-F. Moens, "Cross-language linking of news stories on the Web using interlingual topic modeling," in *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining (SWSM@CIKM)*, 2009, pp. 57–64.

[68] W. De Smet, J. Tang, and M.-F. Moens, "Knowledge transfer across multilingual corpora via latent topics," in *Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2011, pp. 549–560.

[69] H. Déjean, E. Gaussier, and F. Sadat, "An approach based on multilingual thesauri and model combination for bilingual lexicon extraction," in *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, 2002, pp. 1–7.

[70] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[71] K. Deschacht, J. De Belder, and M.-F. Moens, "The latent words language model," *Computer Speech & Language*, vol. 26, no. 5, pp. 384–409, 2012.

[72] R. Deveaud, E. SanJuan, and P. Bellot, "Are semantically coherent topic models useful for ad hoc information retrieval?" in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013, pp. 148–152.

[73] M. T. Diab and S. Finch, "A statistical translation model using comparable corpora," in *Proceedings of the 6th Triennial Conference on Recherche d'Information Assistée par Ordinateur (RIAO)*, 2000, pp. 1500–1508.

[74] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[75] C. H. Q. Ding, T. Li, and W. Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing," *Computational Statistics & Data Analysis*, vol. 52, no. 8, pp. 3913–3927, 2008.

[76] G. Dinu and M. Lapata, "Measuring distributional similarity in context," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010, pp. 1162–1172.

[77] G. Dinu and M. Lapata, "Topic models for meaning similarity in context," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 250–258.

[78] G. Dinu, S. Thater, and S. Laue, "A comparison of models of word meaning in context," in *Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2012, pp. 611–615.

[79] L. Dolamić and J. Savoy, "Retrieval effectiveness of machine translated queries," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 11, pp. 2266–2273, 2010.

[80] B. J. Dorr, G.-A. Levow, and D. Lin, "Construction of a Chinese-English verb lexicon for machine translation and embedded multilingual applications," *Machine Translation*, vol. 17, no. 2, pp. 99–137, 2002.

[81] S. T. Dumais, T. K. Landauer, and M. Littman, "Automatic cross-linguistic information retrieval using Latent Semantic Indexing," in *Proceedings of the SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996, pp. 16–23.

[82] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[83] G. Durrett, A. Pauls, and D. Klein, "Syntactic transfer using a bilingual lexicon," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 1–11.

[84] K. Erk and S. Padó, "A structured vector space model for word meaning in context," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008, pp. 897–906.

[85] K. Erk and S. Padó, "Exemplar-based models for word meaning in context," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 92–97.

[86] R. Feldman and J. Sanger, *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press, 2007.

[87] Y. Feng and M. Lapata, "Visual information in semantic representation," in *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2010, pp. 91–99.

[88] E. Fernandez-Ordonez, R. Mihalcea, and S. Hassan, "Unsupervised word sense disambiguation with multilingual representations," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 847–851.

[89] D. Fišer and N. Ljubešić, "Bilingual lexicon extraction from comparable corpora for closely related languages," in *Proceedings of the 2nd Conference on Recent Advances in Natural Language Processing (RANLP)*, 2011, pp. 125–131.

[90] C. Fox and S. Roberts, "A tutorial on variational Bayesian inference," *Artificial Intelligence Review*, vol. 38, no. 2, pp. 85–95, 2012.

[91] K. Fukumasu, K. Eguchi, and E. P. Xing, "Symmetric correspondence topic models for multilingual text analysis," in *Procedings of the 25th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1295–1303.

[92] P. Fung and P. Cheung, "Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004, pp. 57–63.

[93] P. Fung and L. Y. Yee, "An IR approach for translating new words from nonparallel, comparable texts," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (ACL-COLING)*, 1998, pp. 414–420.

[94] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007, pp. 1606–1611.

[95] E. Gabrilovich and S. Markovitch, "Harnessing the expertise of 70, 000 human editors: Knowledge-based feature generation for text categorization," *Journal of Machine Learning Research*, vol. 8, pp. 2297–2345, 2007.

[96] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 75–102, 1993.

[97] K. Ganchev and D. Das, "Cross-lingual discriminative learning of sequence models with posterior regularization," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1996–2006.

[98] D. Ganguly, J. Leveling, and G. Jones, "Cross-lingual topical relevance models," in *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, 2012, pp. 927–942.

[99] N. Garera, C. Callison-Burch, and D. Yarowsky, "Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences," in *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL)*, 2009, pp. 129–137.

[100] É. Gaussier, "Flow network models for word alignment and terminology extraction from bilingual corpora," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING)*, 1998, pp. 444–450.

[101] É. Gaussier and C. Goutte, "Relation between PLSA and NMF and implications," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2005, pp. 601–602.

[102] É. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte, and H. Déjean, "A geometric view on bilingual lexicon extraction from comparable corpora," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004, pp. 526–533.

[103] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[104] F. C. Gey, H. Jiang, A. Chen, and R. R. Larson, "Manual queries and machine translation in cross-language retrieval and interactive retrieval with Cheshire II at TREC-7," in *Proceedings of the 7th Text Retrieval Conference (TREC)*, 1998, pp. 463–476.

[105] M. Girolami and A. Kabán, "On an equivalence between PLSI and LDA," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR)*, 2003, pp. 433–434.

[106] A. M. Gliozzo, M. Pennacchiotti, and P. Pantel, "The domain restriction hypothesis: Relating term similarity and semantic consistency," in *Proceedings of the 9th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2007, pp. 131–138.

[107] T. Gollins and M. Sanderson, "Improving cross language information retrieval with triangulated translation," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2001, pp. 90–95.

[108] J. T. Goodman, "A bit of progress in language modeling," *Computer Speech & Language*, vol. 15, no. 4, pp. 403–434, 2001.

[109] M. R. Gormley, M. Dredze, B. V. Durme, and J. Eisner, "Shared components topic models," in *Proceedings of the Annual Conference of the North-American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2012, pp. 783–792.

[110] E. Grefenstette and M. Sadrzadeh, "Experimental support for a categorical compositional distributional model of meaning," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011, pp. 1394–1404.

[111] G. Grefenstette, *Cross-Language Information Retrieval*. Kluwer Academic Publishing, 1998.

[112] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 2004, pp. 5228–5235.

[113] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, "Integrating topics and syntax," in *Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS)*, 2004, pp. 537–544.

[114] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychological Review*, vol. 114, no. 2, pp. 211–244, 2007.

[115] W. Guo and M. Diab, "COLEPL and COLSLM: An unsupervised WSD approach to multilingual lexical substitution, tasks 2 and 3 SemEval 2010," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 129–133.

[116] L. A. Ha, G. Fernandez, R. Mitkov, and G. C. Pastor, "Mutual bilingual terminology extraction," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, 2008, pp. 1818–1824.

[117] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein, "Learning bilingual lexicons from monolingual corpora," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2008, pp. 771–779.

[118] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008, pp. 363–371.

[119] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 23, pp. 146–162, 1954.

[120] S. Hassan, C. Banea, and R. Mihalcea, "Measuring semantic relatedness using multilingual representations," in *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM)*, 2012, pp. 20–29.

[121] S. Hassan and R. Mihalcea, "Cross-lingual semantic relatedness using encyclopedic knowledge," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009, pp. 1192–1201.

[122] S. Hassan and R. Mihalcea, "Semantic relatedness using salient semantic analysis," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*, 2011, pp. 884–889.

[123] A. Hazem and E. Morin, "Adaptive dictionary for bilingual lexicon extraction from comparable corpora," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 288–292.

[124] T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, and K. Järvelin, "Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002," *Information Retrieval*, vol. 7, no. 1-2, pp. 99–119, 2004.

[125] G. Heinrich, "Parameter estimation for text analysis," Tech. Rep., 2008.

[126] D. Hiemstra, "A linguistically motivated probabilistic model of information retrieval," in *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 1998, pp. 569–584.

[127] D. Hiemstra and F. de Jong, "Disambiguation strategies for cross-language information retrieval," in *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 1999, pp. 274–293.

[128] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999, pp. 289–296.

[129] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999, pp. 50–57.

[130] D. Hu and L. K. Saul, "A probabilistic topic model for unsupervised learning of musical key-profiles," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 441–446.

[131] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012, pp. 873–882.

[132] R. Huang and E. Riloff, "Bootstrapped training of event extraction classifiers," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2012, pp. 286–295.

[133] D. A. Hull and G. Grefenstette, "Querying across languages: A dictionary-based approach to multilingual information retrieval," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1996, pp. 49–57.

[134] A. Ismail and S. Manandhar, "Bilingual lexicon extraction from comparable corpora using in-domain terms," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 481–489.

[135] J. Jagarlamudi and H. Daumé III, "Extracting multilingual topics from unaligned comparable corpora," in *Proceedings of the 32nd Annual European Conference on Advances in Information Retrieval (ECIR)*, 2010, pp. 444–456.

[136] J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2012, pp. 204–213.

[137] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[138] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Transactions on Information Theory*, vol. 21, no. 3, pp. 250–256, 1975.

[139] W. P. Jones and G. W. Furnas, "Pictures of relevance: A geometric analysis of similarity measures," *Journal of the American Society for Information Science*, vol. 38, no. 6, pp. 420–442, 1987.

[140] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[141] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, 2000.

[142] D. Kartsaklis and M. Sadrzadeh, "Prior disambiguation of word tensors for constructing sentence vectors," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1590–1601.

[143] M. Kay and M. Röscheisen, "Text-translation alignment," *Computational Linguistics*, vol. 19, no. 1, pp. 121–142, 1993.

[144] J. Kazama, S. D. Saeger, K. Kuroda, M. Murata, and K. Torisawa, "A Bayesian method for robust estimation of distributional similarities," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 247–256.

[145] M. D. Kernighan, K. W. Church, and W. A. Gale, "A spelling correction program based on a noisy channel model," in *Proceedings of the 13th International Conference on Computational Linguistics (COLING)*, 1990, pp. 205–210.

[146] D. Kiela and S. Clark, "Detecting compositionality of multi-word expressions using nearest neighbours in vector space models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1427–1432.

[147] A. Kilgarriff, "Comparing corpora," *International Journal of Corpus Linguistics*, vol. 6, no. 1, pp. 1–37, 2001.

[148] S. Kim, K. Toutanova, and H. Yu, "Multilingual named entity recognition using parallel data and metadata from Wikipedia," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012, pp. 694–702.

[149] K. Kishida and E. Ishita, "Translation disambiguation for cross-language information retrieval using context-based translation probability," *Journal of Information Science*, vol. 35, no. 4, pp. 481–495, 2009.

[150] A. L. Kjær, "Phrasemes in legal texts," in *Phraseology: An International Handbook of Contemporary Research*. WdG, 2007, pp. 506–516.

[151] D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," in *Proceedings of the 15th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2002, pp. 3–10.

[152] A. Klementiev, A. Irvine, C. Callison-Burch, and D. Yarowsky, "Toward statistical machine translation without parallel corpora," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2012, pp. 130–140.

[153] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the 10th Machine Translation Summit (MT SUMMIT)*, 2005, pp. 79–86.

[154] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2010.

[155] P. Koehn and K. Knight, "Learning a translation lexicon from monolingual corpora," in *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition (ULA)*, 2002, pp. 9–16.

[156] D. Koller and N. Friedman, *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

[157] Z. Kozareva and E. H. Hovy, "Not all seeds are equal: Measuring the quality of text mining seeds," in *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2010, pp. 618–626.

[158] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Proceedings of the 21nd Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 897–904.

[159] T. K. Landauer and S. T. Dumais, "Solutions to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.

[160] A. Laroche and P. Langlais, "Revisiting context-based projection methods for term-translation spotting in comparable corpora," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 617–625.

[161] V. Lavrenko and J. Allan, "Real-time query expansion in relevance models," CIIR Technical Report IR-473, Tech. Rep., 2006.

[162] V. Lavrenko, M. Choquette, and W. B. Croft, "Cross-lingual relevance models," in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2002, pp. 175–182.

[163] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2001, pp. 120–127.

[164] F. Laws, L. Michelbacher, B. Dorow, C. Scheible, U. Heid, and H. Schütze, "A linguistically grounded graph model for bilingual lexicon extraction," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 614–622.

[165] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the 12th Conference on Advances in Neural Information Processing Systems (NIPS)*, 1999, pp. 556–562.

[166] J. H. Lee, A. Reneaer, and L. C. Smith, "Known-item search: Variations on a concept," in *Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, 2006.

[167] L. Lee, "Measures of distributional similarity," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1999, pp. 25–32.

[168] E. Lefever and V. Hoste, "Construction of a benchmark data set for cross-lingual word sense disambiguation," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, 2010, pp. 1584–1590.

[169] E. Lefever and V. Hoste, "SemEval-2010 task 3: Cross-lingual word sense disambiguation," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 15–20.

[170] E. Lefever and V. Hoste, "SemEval-2013 task 10: Cross-lingual word sense disambiguation," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SEMEVAL)*, 2013, pp. 158–166.

[171] E. Lefever, L. Macken, and V. Hoste, "Language-independent bilingual terminology extraction from a multilingual parallel corpus," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2009, pp. 496–504.

[172] G.-A. Levow, D. W. Oard, and P. Resnik, "Dictionary-based techniques for cross-language information retrieval," *Information Processing and Management*, vol. 41, no. 3, pp. 523–547, 2005.

[173] B. Li and É. Gaussier, "Improving corpus comparability for bilingual lexicon extraction from comparable corpora," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, 2010, pp. 644–652.

[174] B. Li, É. Gaussier, and A. Aizawa, "Clustering comparable corpora for bilingual lexicon extraction," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011, pp. 473–478.

[175] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 524–531.

[176] W. Li, D. M. Blei, and A. McCallum, "Nonparametric Bayes Pachinko allocation," in *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007, pp. 243–250.

[177] D. Lin, "Automatic retrieval and clustering of similar words," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics(ACL-COLING)*, 1998, pp. 768–774.

[178] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[179] M. Littman, S. T. Dumais, and T. K. Landauer, "Automatic cross-language information retrieval using Latent Semantic Indexing," in *Chapter 5 of Cross-Language Information Retrieval*. Kluwer Academic Publishers, 1998, pp. 51–62.

[180] X. Liu, K. Duh, and Y. Matsumoto, "Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus," in *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL)*, 2013, pp. 212–221.

[181] B. Lu and B. K. Tsou, "Towards bilingual term extraction in comparable patents," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2009, pp. 755–762.

[182] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA," *Information Retrieval*, vol. 14, no. 2, pp. 178–203, 2011.

[183] Y. Luo, Z. Le, and M. Wang, "Cross-lingual information retrieval model based on bilingual topic correlation," *Journal of Computational Information Systems*, vol. 9, no. 6, pp. 2433–2440, 2013.

[184] L. Mahapatra, M. Mohan, M. Khapra, and P. Bhattacharyya, "OWNS: Cross-lingual word sense disambiguation using weighted overlap counts and WordNet based similarity measures," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 138–141.

[185] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[186] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[187] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[188] B. Mathieu, R. Besançon, and C. Fluhr, "Multilingual document clusters discovery," in *Proceedings of the 7th Triennial Conference on Recherche d'Information Assistée par Ordinateur (RIAO)*, 2004, pp. 116–125.

[189] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 1, pp. 157–169, 2004.

[190] A. Mauser, S. Hasan, and H. Ney, "Extending statistical machine translation with discriminative and trigger-based lexicon models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009, pp. 210–218.

[191] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on Enron and academic e-mail," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 249–272, 2007.

[192] D. McCarthy and R. Navigli, "SemEval-2007 task 10: English lexical substitution task," in *Proceedings of the 4th International Workshop on Semantic Evaluation (SEMEVAL)*, 2007, pp. 48–53.

[193] J. P. McCrae, P. Cimiano, and R. Klinger, "Orthonormal explicit topic analysis for cross-lingual document matching," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1732–1740.

[194] T. McIntosh and J. R. Curran, "Reducing semantic drift with bagging and distributional similarity," in *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2009, pp. 396–404.

[195] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proceedings of the 17th International Conference on World Wide Web (WWW)*, 2008, pp. 101–110.

[196] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proceedings of the 16th International Conference on World Wide Web (WWW)*, 2007, pp. 171–180.

[197] I. D. Melamed, "Models of translational equivalence among words," *Computational Linguistics*, vol. 26, no. 2, pp. 221–249, 2000.

[198] P. Merlo, S. Stevenson, V. Tsang, and G. Allaria, "A multilingual paradigm for automatic verb classification," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 207–214.

[199] R. Mihalcea and T. Pedersen, "An evaluation exercise for word alignment," in *Proceedings of the NAACL-HLT 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 2003, pp. 1–10.

[200] R. Mihalcea, R. Sinha, and D. McCarthy, "SemEval-2010 task 2: Cross-lingual lexical substitution," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 9–14.

[201] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with Pachinko allocation," in *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007, pp. 633–640.

[202] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression," in *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI)*, 2008, pp. 411–418.

[203] D. Mimno, H. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, "Polylingual topic models," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009, pp. 880–889.

[204] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, July 2011, pp. 262–272.

[205] T. P. Minka and J. D. Lafferty, "Expectation-propagation for the generative aspect model," in *Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence (UAI)*, 2002, pp. 352–359.

[206] J. Mitchell and M. Lapata, "Vector-based models of semantic composition," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008, pp. 236–244.

[207] J. Mitchell and M. Lapata, "Composition in distributional models of semantics," *Cognitive Science*, vol. 34, no. 8, pp. 1388–1429, 2010.

[208] M.-F. Moens, J. Li, and T.-S. Chua, Eds., *Mining User Generated Content*. Chapman and Hall/CRC, 2014.

[209] R. C. Moore, "Towards a simple and accurate statistical approach to learning translation relationships among words," in *Proceedings of the Workshop on Data-Driven Methods in Machine Translation*, 2001, pp. 1–8.

[210] R. C. Moore, "Fast and accurate sentence alignment of bilingual corpora," in *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users (AMTA)*, 2002, pp. 135–144.

[211] R. C. Moore, "Improving IBM word-alignment Model 1," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004, pp. 518–525.

[212] R. C. Moore, "On log-likelihood-ratios and the significance of rare events," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004, pp. 333–340.

[213] E. Morin, B. Daille, K. Takeuchi, and K. Kageura, "Bilingual terminology mining - using brain, not brawn comparable corpora," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007, pp. 664–671.

[214] D. S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Computational Linguistics*, vol. 31, no. 4, pp. 477–504, 2005.

[215] D. S. Munteanu and D. Marcu, "Extracting parallel sub-sentential fragments from non-parallel corpora," in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and the 21st International Conference on Computational Linguistics (ACL-COLING)*, 2006, pp. 81–88.

[216] T. Muramatsu and T. Mori, "Integration of pLSA into probabilistic CLIR model," in *Proceedings of the 4th NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization (NTCIR)*, 2004.

[217] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[218] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

[219] D. L. Nelson, C. L. McEvoy, and S. Dennis, "What is free association and what does it measure?" *Memory and Cognition*, vol. 28, pp. 887–899, 2000.

[220] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proceedings of the 10th Conference of the North-American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2010, pp. 100–108.

[221] X. Ni, J.-T. Sun, J. Hu, and Z. Chen, "Mining multilingual topics from Wikipedia," in *Proceedings of the 18th International World Wide Web Conference (WWW)*, 2009, pp. 1155–1156.

[222] X. Ni, J.-T. Sun, J. Hu, and Z. Chen, "Cross lingual text classification by mining multilingual topics from Wikipedia," in *Proceedings of the 4th International Conference on Web Search and Web Data Mining (WSDM)*, 2011, pp. 375–384.

[223] J.-Y. Nie, *Cross-Language Information Retrieval*. Morgan & Claypool Publishers, 2010.

[224] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999, pp. 74–81.

[225] Y. Niwa and Y. Nitta, "Co-occurrence vectors from corpora vs. distance vectors from dictionaries," in *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, 1994, pp. 304–309.

[226] P. Niyogi and N. Karmarkar, "An approach to data reduction and clustering with theoretical guarantees," in *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2000, pp. 679–686.

[227] D. Ó Séaghdha, "Latent variable models of selectional preference," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 435–444.

[228] D. Ó Séaghdha and A. Korhonen, "Probabilistic models of similarity in syntactic context," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011, pp. 1047–1057.

[229] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[230] P. Ogilvie and J. Callan, "Combining document representations for known-item search," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2003, pp. 143–150.

[231] S. Padó and M. Lapata, "Dependency-based construction of semantic space models," *Computational Linguistics*, vol. 33, no. 2, pp. 161–199, 2007.

[232] S. Padó and M. Lapata, "Cross-lingual annotation projection for semantic roles," *Journal of Artificial Intelligence Research*, vol. 36, pp. 307–340, 2009.

[233] P. Pantel and D. Lin, "Discovering word senses from text," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002, pp. 613–619.

[234] Y. Peirsman and S. Padó, "Cross-lingual induction of selectional preferences with bilingual vector spaces," in *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2010, pp. 921–929.

[235] Y. Peirsman and S. Padó, "Semantic relations in bilingual lexicons," *ACM Transactions on Speech and Language Processing*, vol. 8, no. 2, p. article 3, 2011.

[236] V. Pekar, R. Mitkov, D. Blagoev, and A. Mulloni, "Finding translations for low-frequency words in comparable corpora," *Machine Translation*, vol. 20, no. 4, pp. 247–266, 2006.

[237] C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, Eds., *Evaluation of Cross-Language Information Retrieval Systems, 2nd Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Revised Papers*, 2002.

[238] C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, Eds., *Advances in Cross-Language Information Retrieval, 3rd Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Revised Papers*, 2003.

[239] S. Petrov, D. Das, and R. T. McDonald, "A universal part-of-speech tagset," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 2089–2096.

[240] D. Pinto, J. Civera, A. Barrón-Cedeòo, A. Juan, and P. Rosso, "A statistical approach to crosslingual natural language tasks," *Journal of Algorithms*, vol. 64, no. 1, pp. 51–60, 2009.

[241] J. C. Platt, K. Toutanova, and W.-T. Yih, "Translingual document representations from discriminative projections," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010, pp. 251–261.

[242] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1998, pp. 275–281.

[243] E. Prochasson and P. Fung, "Rare word translation extraction from aligned comparable documents," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011, pp. 1327–1335.

[244] M. Purver, K. Körding, T. Griffiths, and J. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and the 21st International Conference on Computational Linguistics (ACL-COLING)*, 2006, pp. 17–24.

[245] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge University Press, 2011.

[246] R. Rapp, "Identifying word translations in non-parallel texts," in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1995, pp. 320–322.

[247] R. Rapp, "Automatic identification of word translations from unrelated English and German corpora," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1999, pp. 519–526.

[248] S. Reddy, I. P. Klapaftis, D. McCarthy, and S. Manandhar, "Dynamic and static prototype vectors for semantic composition," in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, 2011, pp. 705–713.

[249] U. Reinke, "State of the art in translation memory technology," *Translation: Computation, Corpora, Cognition*, vol. 3, no. 1, 2013.

[250] J. Reisinger and R. J. Mooney, "A mixture model with sharing for lexical semantics," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2010, pp. 1173–1182.

[251] J. Reisinger and R. J. Mooney, "Multi-prototype vector-space models of word meaning," in *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2010, pp. 109–117.

[252] P. Resnik and N. A. Smith, "The Web as a parallel corpus," *Computational Linguistics*, vol. 29, no. 3, pp. 349–380, 2003.

[253] J. Richardson, T. Nakazawa, and S. Kurohashi, "Robust transliteration mining from comparable corpora with bilingual topic models," in *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, 2013, pp. 261–269.

[254] E. Riloff and J. Shepherd, "A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction," *Natural Language Engineering*, vol. 5, no. 2, pp. 147–156, 1999.

[255] A. Ritter, Mausam, and O. Etzioni, "A latent Dirichlet allocation method for selectional preferences," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 424–434.

[256] S. E. Robertson, "The probabilistic character of relevance," *Information Processing and Management*, vol. 13, no. 4, pp. 247–251, 1977.

[257] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter, "Probabilistic models of indexing and searching," in *Proceedings of the 3rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1980, pp. 35–56.

[258] S. E. Robertson and S. Walker, "On relevance weights with little relevance information," in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1997, pp. 16–24.

[259] S. Roller and S. Schulte im Walde, "A multimodal LDA model integrating textual, cognitive and visual modalities," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1146–1157.

[260] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[261] B. Roth and D. Klakow, "Combining Wikipedia-based concept models for cross-language retrieval," in *Proceedings of the 1st Information Retrieval Facility Conference (IRFC)*, 2010, pp. 47–59.

[262] A. Rudnick, C. Liu, and M. Gasser, "HLTDI: CL-WSD using Markov random fields for SemEval-2013 task 10," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SEMEVAL)*, 2013, pp. 171–177.

[263] S. Rudolph and E. Giesbrecht, "Compositional matrix-space models of language," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 907–916.

[264] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1605–1614.

[265] S. J. Russell and P. Norvig, *Artificial Intelligence - A Modern Approach, 3rd Edition*. Pearson Education, 2010.

[266] J. Savoy, "Combining multiple strategies for effective monolingual and cross-language retrieval," *Information Retrieval*, vol. 7, no. 1-2, pp. 121–148, 2004.

[267] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the International Conference on New Methods in Language Processing*, 1994.

[268] P. Sheridan and J. P. Ballerini, "Experiments in multilingual information retrieval using the SPIDER system," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1996, pp. 58–65.

[269] D. Shezaf and A. Rappoport, "Bilingual lexicon generation using non-aligned signatures," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 98–107.

[270] C. Silberer and S. P. Ponzetto, "UHD: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 134–137.

[271] D. A. Smith and J. Eisner, "Bootstrapping feature-rich dependency parsers with entropic priors," in *Proceedings of the Joint 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 667–677.

[272] P. Smolensky, "Tensor product variable binding and the representation of symbolic structures in connectionist systems," *Artificial Intelligence*, vol. 46, no. 1-2, pp. 159–216, 1990.

[273] B. Snyder and R. Barzilay, "Climbing the tower of Babel: Unsupervised multilingual learning," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010, pp. 29–36.

[274] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proceedings of the 24th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 801–809.

[275] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 1201–1211.

[276] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 2142–2147.

[277] K. Stevens, W. P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 952–961.

[278] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of Latent Semantic Analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[279] M. Steyvers, R. M. Shiffrin, and D. L. Nelson, "Word association spaces for predicting semantic similarity effects in episodic memory," in *Experimental Cognitive Psychology and Its Applications*, 2004, pp. 237–249.

[280] O. Täckström, D. Das, S. Petrov, R. McDonald, and J. Nivre, "Token and type constraints for cross-lingual part-of-speech tagging," *Transactions of ACL*, vol. 1, pp. 1–12, 2013.

[281] O. Täckström, R. McDonald, and J. Nivre, "Target language adaptation of discriminative transfer parsers," in *Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2013, pp. 1061–1071.

[282] A. Takasu, "Cross-lingual keyword recommendation using latent topics," in *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, 2010, pp. 52–56.

[283] A. Tamura, T. Watanabe, and E. Sumita, "Bilingual lexicon extraction from comparable corpora using label propagation," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 24–36.

[284] T. Tao and C. Zhai, "Mining comparable bilingual text corpora for cross-language information integration," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2005, pp. 691–696.

[285] A. Tchechmedjiev, J. Goulian, and D. Schwab, "Fusion strategies applied to multilingual features for an knowledge-based word sense disambiguation algorithm: Evaluation and comparison," in *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, 2013, pp. 67–78.

[286] S. Thater, H. Fürstenau, and M. Pinkal, "Contextualizing semantic representations using syntactically enriched vector models," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 948–957.

[287] S. Thater, H. Fürstenau, and M. Pinkal, "Word meaning in context: A simple and effective vector model," in *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, 2011, pp. 1134–1143.

[288] M. Thelen and E. Riloff, "A bootstrapping method for learning semantic lexicons using extraction pattern contexts," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 214–221.

[289] J. Tiedemann, "Combining clues for word alignment," in *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2003, pp. 339–346.

[290] J. Tiedemann, "News from OPUS - A collection of multilingual parallel corpora with tools and interfaces," in *Proceedings of the 1st Conference on Recent Advances in Natural Language Processing (RANLP)*, 2009, pp. 237–248.

[291] J. Tiedemann, *Bitext Alignment*. Morgan & Claypool Publishers, 2011.

[292] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[293] I. Titov and R. T. McDonald, "A joint model of text and aspect ratings for sentiment summarization," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2008, pp. 308–316.

[294] K. Tu and V. Honavar, "Unambiguity regularization for unsupervised learning of probabilistic grammars," in *Proceedings of the Joint 2012 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012, pp. 1324–1334.

[295] D. Turcato, "Automatically creating bilingual lexicons for machine translation from bilingual text," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1998, pp. 1299–1306.

[296] F. Ture, J. Lin, and D. Oard, "Combining statistical translation techniques for cross-language information retrieval," in *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, 2012, pp. 2685–2702.

[297] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[298] P. D. Turney, "Similarity of semantic relations," *Computational Linguistics*, vol. 32, no. 3, pp. 379–416, 2006.

[299] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artifical Intelligence Research*, vol. 37, no. 1, pp. 141–188, 2010.

[300] H. R. Turtle and W. B. Croft, "Efficient probabilistic inference for text retrieval," in *Proceedings of the 3rd Triennial Conference on Recherche d'Information Assistée par Ordinateur (RIAO)*, 1991, pp. 644–662.

[301] N. Ueffing and H. Ney, "Word-level confidence estimation for machine translation," *Computational Linguistics*, vol. 33, no. 1, pp. 9–40, 2007.

[302] M. Utiyama and H. Isahara, "Reliable measures for aligning Japanese-English news articles and sentences," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003, pp. 72–79.

[303] T. Van de Cruys, T. Poibeau, and A. Korhonen, "Latent vector weighting for word meaning in context," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011, pp. 1012–1022.

[304] L. van der Plas, P. Merlo, and J. Henderson, "Scaling up automatic cross-lingual semantic role annotation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011, pp. 299–304.

[305] M. van Gompel and A. van den Bosch, "WSD2: Parameter optimisation for memory-based cross-lingual word-sense disambiguation," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SEMEVAL)*, 2013, pp. 183–187.

[306] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979.

[307] T. Vanallemeersch, "Belgisch Staatsblad corpus: Retrieving French-Dutch sentences from official documents," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, 2010, pp. 3413–3416.

[308] A. Venugopal, S. Vogel, and A. Waibel, "Effective phrase translation extraction from alignment models," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, 2003, pp. 319–326.

[309] E. M. Voorhees, "The TREC-8 question answering track report," in *Proceedings of the 8th Text Retrieval Conference (TREC)*, 1999.

[310] P. Vossen, Ed., *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1998.

[311] T. Vu, A. T. Aw, and M. Zhang, "Feature-based method for document alignment in comparable news corpora," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2009, pp. 843–851.

[312] I. Vulić, W. De Smet, and M.-F. Moens, "Identifying word translations from comparable corpora using latent topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011, pp. 479–484.

[313] I. Vulić, W. De Smet, and M.-F. Moens, "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," *Information Retrieval*, vol. 16, no. 3, pp. 331–368, 2013.

[314] I. Vulić and M.-F. Moens, "Cross-lingual semantic similarity of words as the similarity of their semantic word responses," in *Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2013, pp. 106–116.

[315] I. Vulić and M.-F. Moens, "A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else)," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pp. 1613–1624.

[316] I. Vulić, S. Zoghbi, and M.-F. Moens, "Learning to bridge colloquial and formal language applied to linking and search of e-commerce data," in *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), to appear*, 2014, pp. xx–xx.

[317] H. Wallach, "Topic modeling: Beyond bag-of-words," in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 977–984.

[318] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, 2009, pp. 1105–1112.

[319] J. Wang and D. W. Oard, "Combining bidirectional translation and synonymy for cross-language information retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2006, pp. 202–209.

[320] X. Wang and E. Grimson, "Spatial Latent Dirichlet Allocation," in *Proceedings of the 20th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2007.

[321] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539–555, 2009.

[322] X. Wang and A. McCallum, "Topics over time: A non-Markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (CIKM)*, 2006, pp. 424–433.

[323] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2006, pp. 178–185.

[324] R. M. Weischedel, M. Meteer, R. M. Schwartz, L. A. Ramshaw, and J. Palmucci, "Coping with ambiguity and unknown words through probabilistic models," *Computational Linguistics*, vol. 19, no. 2, pp. 359–382, 1993.

[325] R. Wicentowski, M. Kelly, and R. Lee, "SWAT: Cross-lingual lexical substitution using local context matching, bilingual dictionaries and machine translation," in *Proceedings of the 5th International Workshop on Semantic Evaluation (SEMEVAL)*, 2010, pp. 123–128.

[326] H. Wu, H. Wang, and C. Zong, "Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, 2008, pp. 993–1000.

[327] J. Xu, R. Weischedel, and C. Nguyen, "Evaluating a probabilistic model for cross-lingual information retrieval," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2001, pp. 105–110.

[328] G.-R. Xue, W. Dai, Q. Yang, and Y. Yu, "Topic-bridged pLSA for cross-domain text classification," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2008, pp. 627–634.

[329] D. Yarowsky and G. Ngai, "Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora," in *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2001, pp. 200–207.

[330] X. Yi and J. Allan, "A comparative study of utilizing topic models for information retrieval," in *Proceedings of the 31th European Conference on Advances in Information Retrieval (ECIR)*, 2009, pp. 29–41.

[331] E. Zavitsanos, G. Paliouras, and G. A. Vouros, "Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes," *Journal of Machine Learning Research*, vol. 12, pp. 2749–2775, 2011.

[332] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems*, vol. 22, no. 2, pp. 179–214, 2004.

[333] D. Zhang, Q. Mei, and C. Zhai, "Cross-lingual latent topic extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010, pp. 1128–1137.

[334] T. Zhang, K. Liu, and J. Zhao, "Cross lingual entity linking with bilingual topic model," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 2218–2224.

[335] B. Zhao and E. P. Xing, "BiTAM: Bilingual topic admixture models for word alignment," in *Proceedings of the 44th Annual Meeting of the Association for Computational Linguisticsthe and the 21st International Conference on Computational Linguistics and (ACL-COLING)*, 2006, pp. 969–976.

[336] B. Zhao and E. P. Xing, "HM-BiTAM: Bilingual topic exploration, word alignment, and translation," in *Proceedings of the 21st Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 1689–1696.

[337] H. Zhao, Y. Song, C. Kit, and G. Zhou, "Cross language dependency parsing using a bilingual lexicon," in *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2009, pp. 55–63.

[338] Z. Zhu, M. Li, L. Chen, and Z. Yang, "Building comparable corpora based on bilingual LDA model," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2013, pp. 278–282.

[339] G. K. Zipf, *The Psychobiology of Language*. Houghton-Mifflin, 1935.

[340] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

[341] S. Zoghbi, I. Vulić, and M.-F. Moens, "Are words enough? A study on text-based representations and retrieval models for linking pins to online shops," in *Proceedings of the 2013 International CIKM Workshop on Mining Unstructured Big Data Using Natural Language Processing (UnstructureNLP@CIKM 2013)*, 2013, pp. 45–52.

[342] S. Zoghbi, I. Vulić, and M.-F. Moens, "I pinned it. Where can I buy one like it? Automatically linking Pinterest pins to online webshops," in *Proceedings of the 2013 International CIKM Workshop on Data-driven User Behavioral Modeling and Mining from Social Media (DUBMOD@CIKM 2013)*, 2013, pp. 9–12.

# Curriculum Vitae

Ivan Vulić was born in Zadar, Republic of Croatia on May 7th, 1986. He received the degree of Master in Computer Science from the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia in September 2009. During his studies, he was awarded four annual Dean's awards, and the bronze plaque "Josip Lončar" as the best graduated student in his class. In November 2009, he joined the LIIR (Language Intelligence & Information Retrieval) research group at the Department of Computer Science, KU Leuven, Belgium as a predoctoral student. In April 2010, he started his Ph.D. program with the emphasis on models of multilingual text mining and their applications in natural language processing and information retrieval.

Since 2011 he serves as an elected member of the student board of the European Chapter of the Association for Computational Linguistics (EACL). He served as a co-chair of the Student Research Workshop at the EACL 2014 conference held in April 2014 in Gothenburg, Sweden.

# List of Publications

## Journal Articles

[1] **I. Vulić** and M.-F. Moens. "A probabilistic framework for modeling cross-lingual semantic similarity in context based on latent cross-lingual concepts," journal article in preparation for *Journal of Artificial Intelligence Research*, 2014.

[2] S. Zoghbi, **I. Vulić**, and M.-F. Moens. "Cross-idiomatic linking between Web sources with applications in e-commerce," submitted to *Information Retrieval*, 2014.

[3] **I. Vulić**, W. De Smet, J. Tang, and M.-F. Moens. "Probabilistic topic modeling in multilingal settings: An overview of its methodology with applications," accepted with minor revisions in *Information Processing & Management*, 2014.

[4] **I. Vulić**, W. De Smet, and M.-F. Moens. "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora," *Information Retrieval*, vol. 16, no. 3, pp. 331-368, Springer, 2013.

## Peer-Reviewed International Conference Articles

[1] **I. Vulić**, W. De Smet, and M.-F. Moens. "Extracting shared and non-shared topics from non-parallel data with application to cross-lingual knowledge transfer," submitted to the *19th Conference on the Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 25-29 October 2014, pp. xx-xx, ACL, 2014.

[2] **I. Vulić** and M.-F. Moens. "Probabilistic models of cross-lingual semantic similarity based on latent cross-lingual concepts," submitted to the *19th Conference on the Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 25-29 October 2014, pp. xx-xx, ACL, 2014.

[3] **I. Vulić**, S. Zoghbi, and M.-F. Moens. "Learning to bridge colloquial and formal language applied to linking and search of e-commerce data, " accepted for publication in

the *37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Gold Coast, Queensland, Australia, 6-11 July 2014, pp. xx-xx, ACM, 2014.

[4] K. Heylen, S. Bond, D. De Hertog, H. Kockaert, F. Steurs, and **I. Vulić**. "TermWise: Leveraging big data for terminological support in legal translation, " accepted for publication in the *11th International Conference on Terminology and Knowledge Engineering (TKE)*, Berlin, Germany, 19-21 June 2014, pp. xx-xx, Springer, 2014.

[5] K. Heylen, S. Bond, **I. Vulić**, D. De Hertog, and H. Kockaert. "TermWise: A CAT tool with context-sensitive terminological support, " accepted for publication in the *9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 26-31 May 2014, pp. 4018-4022, ELRA, 2014.

[6] M.-F. Moens and **I. Vulić**. "Multilingual probabilistic topic modeling and its applications in Web mining and search," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM)*, New York City, New York, USA, 24-28 February 2014, pp. 681-682, ACM, 2014.

[7] **I. Vulić** and M.-F. Moens. "A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else)," in *Proceedings of the 18th Conference on the Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, Washington, USA, 18-21 October 2013, pp. 1613-1624, ACL, 2013.

[8] N. Shrestha, **I. Vulić**, and M.-F. Moens. "An IR-inspired approach to recovering named entity tags in broadcast news," in *Proceedings of the 6th Information Retrieval Facility Conference (IRFC)*, vol. 8201 of *Lecture Notes in Computer Science*, Limassol, Cyprus, 7-9 October 2013, pp. 45-57, Springer, 2013.

[9] **I. Vulić** and M.-F. Moens. "Cross-lingual semantic similarity of words as the similarity of their semantic word responses," in *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, Georgia, USA, 9-15 June 2013, pp. 106-116, ACL, 2013.

[10] **I. Vulić** and M.-F. Moens. "A unified framework for monolingual and cross-lingual relevance modeling based on probabilistic topic models," in *Proceedings of the 35th European Conference on Information Retrieval (ECIR)*, vol. 7814 of *Lecture Notes in Computer Science*, Moscow, Russian Federation, 24-27 March 2013, pp. 98-109, Springer, 2013.

[11] M.-F. Moens and **I. Vulić**. "Monolingual and cross-lingual probabilistic topic models and their application in information retrieval," in *Proceedings of the 35th European Conference on Information Retrieval (ECIR)*, vol. 7814 of *Lecture Notes in Computer Science*, Moscow, Russian Federation, 24-27 March 2013, pp. 875-878, Springer, 2013.

[12] **I. Vulić** and M.-F. Moens. "Sub-corpora sampling with an application to bilingual lexicon extraction," in *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India, 8-15 December 2012, pp. 2721-2738, ACL, 2012.

[13] **I. Vulić** and M.-F. Moens. "Detecting highly confident word translations from comparable corpora without any prior knowledge," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France, 23-27 April 2012, pp. 449-459, ACL, 2012.

[14] B. Jans, **I. Vulić**, S. Bethard, and M.-F. Moens. "Skip N-grams and ranking functions for predicting script events," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France, 23-27 April 2012, pp. 336-344, ACL, 2012.

[15] **I. Vulić**, W. De Smet, and M.-F. Moens. "Cross-language information retrieval with latent topic models trained on a comparable corpus," in *Proceedings of the 7th Asian Information Retrieval Societies Conference (AIRS)*, vol. 7097 of *Lecture Notes in Computer Science*, Dubai, UAE, 19-21 December 2011, pp. 37-48, Springer, 2011.

[16] **I. Vulić**, W. De Smet and M.-F. Moens. "Identifying word translations from comparable corpora using latent topic models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, Portland, Oregon, USA, 19-24 June 2011, pp. 479-484, ACL, 2011.

## Peer-Reviewed International Workshop Articles

[1] S. Zoghbi, **I. Vulić**, and M.-F. Moens. "Are words enough? A study on text-based representations and retrieval models for linking pins to online shops," in *Proceedings of the CIKM 2013 Workshop on Mining Unstructured Big Data Using Natural Language Processing (MNLP)*, San Francisco, California, USA, 27 October-1 November 2013, pp. 45-52, ACM, 2013.

[2] S. Zoghbi, **I. Vulić**, and M.-F. Moens. "I pinned it. Where can I buy one like it? Automatically linking Pinterest pins to online webshops," in *Proceedings of the CIKM 2013 Workshop on Data-Driven User Behavioral Modelling and Mining from Social Media (DUBMOD)*, San Francisco, California, USA, 27 October-1 November 2013, pp. 9-12, ACM, 2013.

[3] N. Shrestha and **I. Vulić**. "Named entity recognition in broadcast news using similar written texts," in *Proceedings of the Student Research Workshop held in conjunction with the Conference on Recent Advances in Natural Language Processing (RANLP SRW)*, Hissar, Bulgaria, 9-11 September 2013, pp. 142-148, ACL, 2013. (**best student paper)**

[4] **I. Vulić**, W. De Smet, J. Tang, and M.-F. Moens. "Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications," in *NIPS 2012 Workshop on Cross-Lingual Technologies (xLiTe)*, Lake Tahoe, Nevada, USA, 7-8 December 2012, 11 pages, NIPS, 2012.

## Technical Reports

[1] **I. Vulić** and M.-F. Moens "Term alignment: State-of-the-art overview," *Internal Report for the TermWise Project*, 82 pages, Department of Computer Science, KU Leuven, 2010.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
LANGUAGE INTELLIGENCE & INFORMATION RETRIEVAL
Celestijnenlaan 200A box 2402
B-3001 Heverlee
ivan.vulic@cs.kuleuven.be
http://people.cs.kuleuven.be/∼ivan.vulic/