Discovering Missing Wikipedia Inter-language Links by means of Cross-lingual Word Sense Disambiguation

Els Lefever[1,2], **Véronique Hoste**[1,3] **and Martine De Cock**[2]

[1]LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
[2]Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium
[3]Dept. of Linguistics, Ghent University
Blandijnberg 2, 9000 Gent, Belgium
els.lefever, veronique.hoste@hogent.be, martine.decock@ugent.be

Abstract

Wikipedia pages typically contain inter-language links to the corresponding pages in other languages. These links, however, are often incomplete. This paper describes a set of experiments in which the viability of discovering such missing inter-language links for ambiguous nouns by means of a cross-lingual Word Sense Disambiguation approach is investigated. The input for the inter-language link detection system is a set of Dutch pages for a given ambiguous noun and the output of the system is a set of links to the corresponding pages in three target languages (viz. French, Spanish and Italian). The experimental results show that although it is a very challenging task, the system succeeds to detect missing inter-language links between Wikipedia documents for a manually labeled test set. The final goal of the system is to provide a human editor with a list of possible missing links that should be manually verified.

Keywords: Wikipedia links, Cross-lingual WSD, Word Sense Disambiguation

1. Introduction

Wikipedia[1] is a very popular collaborative multilingual web-based encyclopedia that currently contains over 3.7 million articles in English. As Wikipedia is such a huge lexical and knowledge resource, it is a very valuable data source for various NLP tasks such as Word Sense Disambiguation (Mihalcea and Csomai, 2007), Named Entity Disambiguation (Bunescu and Pasca, 2006) or text summarization (Nastase, 2008). In addition, it can also be used for the construction of large monolingual or multilingual resources. A recent example is BabelNet (Navigli and Ponzetto, 2010), a very large multilingual semantic network that combines information extracted from Wikipedia and WordNet (Fellbaum, 1998).

Wikipedia pages typically introduce information about a specific concept and contain hypertext linked to other Wikipedia pages. The title of the page refers to the main concept of the page and its URL consists of a sequence of words separated by underscores. In case the concept is ambiguous, the title also contains disambiguation information between parenthesis. The Dutch noun *bal* for instance has three pages: (1) *Bal_(danspartij)* where *bal* means *ball* in the sense of *party*, (2) *Bal_(voorwerp)* where *bal* means *ball* as an object and (3) *Bal_(wiskunde)* where *bal* refers to *sphere* in a mathematical sense.

In addition, inter-language links are provided to the corresponding pages in other languages. This way, Wikipedia users can consult the relevant information in their mother tongue. These links, however, are not always complete; often a page is only linked to the corresponding page in a limited number of languages. This might be because corresponding pages in other languages are lacking, or, even when they do exist, because no human contributor has established the appropriate inter-language link yet. Our goal is to provide automated support for such cross-lingual link discovery.

The structure of this paper is as follows. In Section 2., we provide a detailed description of the cross-lingual link discovery system. In Section 3., we introduce the scoring metrics that were used for the evaluation of the system and discuss the experimental results. Section 4. concludes this paper and gives some directions for future research.

2. Cross-lingual Link Discovery System

We propose a cross-lingual link discovery system that discovers missing Wikipedia inter-language links to corresponding pages in other languages for ambiguous nouns. Although the framework of our approach is language-independent, we built a system using Dutch as an input language and French, Italian and Spanish as target languages. The input for the system is a set of Dutch pages for a given ambiguous noun (E.g. *Bal*), and the output of the system is a set of links to the corresponding pages in the three target languages. Our link discovery system contains two submodules. In a first step all pages are retrieved that contain a translation (in the three target languages) of the ambiguous word in the page title (*Greedy crawler* module), whereas in a second step all corresponding pages are linked in the focus language (being Dutch in our case) and the three target languages (*Cross-lingual web page linker*).

A closely related domain is monolingual link discovery, which can be used to establish links between Wikipedia pages (Adafre and de Rijke, 2005) or to enrich a text with links to encyclopedic (Wikipedia) knowledge (Mihalcea and Csomai, 2007). More recently, dedicated Cross Language Link Discovery competitions have been orga-

[1] http://www.wikipedia.org

nized where participants have to establish links between the English Wikipedia pages and the corresponding Chinese, Japanese and Korean pages (Huang et al., 2009). The task we present is even more challenging, as we try to discover missing inter-language links for *ambiguous* nouns.

2.1. Greedy crawler

In a first step, we use a crawler to retrieve all Dutch pages containing the ambiguous focus word and all pages in the three target languages that contain a translation of the ambiguous focus word in the page title. In order to find possible translations in our target languages, we use two online dictionaries: the OPUS dictionary[2] that results from running word alignment on a parallel corpus and a Dutch online translation dictionary[3]. In a next step, these translations are embedded in regular expression rules to retrieve all pages from Wikipedia[4] containing these translations in the page title. In a post processing step we filter out pages where the translation is part of a compound or pages that only contain a redirect link to another page. This way we filter for instance *mouse_key* from the English results. Table 1 lists some information on the crawling results for the Dutch word *muis* that refers amongst others to a *computer mouse*, the *animal mouse* or the *ball of the thumb*. Some of the Dutch pages have no corresponding page in the other languages (e.g. *Muis_(klank)* that refers to a *tone ball*[5]), whereas some pages in the other languages (e.g. *Souris_(aéronotique)*[6]) cannot be linked to the Dutch pages for *Muis* because the translation in the other language is polysemous as well.

	Online Translations	Nr of retrieved pages	Nr of pages after filtering
Dutch	muis	6	6
French	souris	159	10
Italian	topo	21	4
	mouse	7	1
	topolino	90	6
Spanish	ratón	180	6
	mouse	17	0
English	mouse	186	5
	ball	130	9
German	Maus	16	2

Table 1: Crawling results for *muis*

In this crawling step, our goal is to have a very high recall and find as much pages as possible. We do not perform any disambiguation at this point.

2.2. Cross-lingual web page linker

The second module compares a Dutch document with all retrieved Wikipedia documents that result from the greedy crawler and determines whether they refer to the same content. This process can be considered as a disambiguation task. Whereas the first step aims to obtain high recall on the retrieved web pages, the web page linker targets high precision on the linked web pages.

In order to solve the disambiguation task, we recasted the linking of two web pages as a classification problem: for every pair of documents, the classifier determines whether they should be linked or not. The framework of our classification approach is adopted from a more general Cross-lingual Word Sense Disambiguation framework (Lefever et al., 2011). The latter approach uses a given input language and word alignment on a parallel corpus to automatically derive word senses (or translations in this case) for target languages. In order to predict a correct translation of an ambiguous noun in one target language, translation features from four other languages are incorporated in the feature vector. The idea to use information from the aligned languages in the parallel corpus starts from the "two languages are more informative than one" hypothesis. Previous research confirmed this hypothesis (Lefever and Hoste, 2011) and even showed that the classification scores increase relatively to the number of languages that is used for adding multilingual evidence to the feature vector.

2.2.1. Training Feature Vectors

To train the classifier we first construct a training corpus from the Europarl parallel corpus (Koehn, 2005). We extract from Europarl all Dutch sentences that contain a given focus word, the preceding and following sentence and the aligned sentences in five other languages (viz. French, Italian, Spanish, German and English). In the case of *muis*, we retrieved 41 occurrences from the Dutch part of the Europarl corpus and compare each Dutch instance to all aligned sentences in the target language. This results in a training corpus of 1681 instances per target language. Each training instance receives a binary classification label; "1" in case the two documents are linked, and "0" in case the documents are not aligned in Europarl. Table 2 lists the number of training instances containing the Dutch ambiguous word and the total size of the original training base.

target word	occurrences	total number of training instances
muis	41	1.681
graad	135	18.225
operatie	775	600.625
stam	68	4.624

Table 2: Distribution and size of the original training base per target language

The training data base that is used to train the three classifiers - one for each target language - is very skewed; for the classification task at hand the number of positive instances is very small compared to the total number of training instances. Previous research has shown that the performance of support vector machines suffers from such

[2] http://opus.lingfil.uu.se/lex.php

[3] www.vertaalwoord.nl

[4] We used the Wikipedia dump files from September 2011 as they can be found on dumps.wikimedia.org

[5] A *tone ball* is a collection of dust inside an instrument that rolls around, getting bigger and more regular over time.

[6] *Souris* used in this sense refers to the inlet cones (sometimes called shock cones or inlet centerbodies) of some supersonic aircraft.

class imbalance in the data set (Raskutti and Kowalczyk, 2003). Therefore, we decided to randomly down-sample the training set in order to have a good distribution of positive and negative examples. In the case of down-sampling, examples from the majority class (negative examples in our case) are removed, whereas in case of up-sampling, examples from the minority class are duplicated. For the three test words *muis, graad* and *stam*, we removed half of the negative training instances. Because of the huge total number of training instances for *operatie* (600,625 instances), we decided to remove in this case 99% of the negative training instances. This way we obtain for each test word a training base that does not exceed 10,000 training instances and that contains between 1.5% and 11% of positive training instances.

The training sentences are linguistically preprocessed by means of the Treetagger (Schmid, 1994) tool that performs tokenization and Part-of-Speech tagging. The preprocessed sentences are used as input to build a set of Bag-of-Words (BoW) features related to the Dutch sentences as well as the aligned sentences in the five other languages. From now on, we will refer to the window of three sentences containing the focus word as a *document*.

For each pair of documents we store three types of BoW features: (1) the content words (nouns, adjectives, adverbs and verbs) from the Dutch document, (2) the content words from the target document and (3) the content words from the documents in the other four languages.

We constructed two flavors of these Bag-of-Words features: a binary version and a weighted version. The first flavor stores a binary label for each content word that indicates whether a word occurs in the document or not. In order to also detect the most relevant keywords for each document, we compute the TF-IDF score (Term Frequency - Inverse Document Frequency) for each word in the given document, i.e.the relative frequency of the word in the document compared to the frequency of the word in the entire document corpus (Salton and Buckley, 1988). This version of the feature vectors stores for every content word a weighted TF-IDF score.

2.2.2. Test Feature Vectors

During the test phase, we try to link Dutch Wikipedia documents to documents in our three target languages. Table 3 gives an overview of the number of considered Wikipedia pages per ambiguous word per language, and the resulting number of test instances and positive links per classifier. In order to identify positive links between test pairs, we have manually inspected and labeled all test pairs ("1" in case there is a link between the two documents, "0" in case the documents should not be linked).

For the creation of the feature vectors for the test documents, we follow a similar strategy as the one we used for the creation of the training instances. The first part of the feature vector contains the Dutch content words, the second part of the feature vector contains the content words from the target document. For the construction of the Bag-of-Words features for the other four languages however, we need to adopt a different approach as we do not have aligned documents at our disposal for new pairs of documents. In order to solve this problem, we use the Google Translate API[7] to automatically generate a translation for the target document in the four other languages that are not the language of the target document.

Dutch-French Wikipedia pages				
	Dutch	French	test pairs	links
muis	6	10	60	8
graad	6	29	174	7
stam	7	24	168	11
operatie	4	7	28	3
Dutch-Spanish Wikipedia pages				
	Dutch	Spanish	test pairs	links
muis	6	6	36	3
graad	6	17	102	4
stam	7	16	112	7
operatie	4	9	36	2
Dutch-Italian Wikipedia pages				
	Dutch	Italian	test pairs	links
muis	6	11	66	2
graad	6	18	108	7
stam	7	9	63	5
operatie	4	10	40	4

Table 3: Number of test pairs and positive links per language combination (Dutch-French, Dutch-Spanish and Dutch-Italian)

2.2.3. Classifier

As a classifier we used the Support Vector Machine (SVM) algorithm implemented in SVMLIGHT (Joachims, 1998). A standard SVM is a supervised learning algorithm for binary classification. Given a training set containing positive and negative examples, an SVM training algorithm maps these examples to a high-dimensional feature space in such a way that the examples of the separate categories are divided by a hyperplane (or decision boundary). New examples are then mapped to the feature space and labeled depending on their position with respect to the decision boundary. The distance between the example and the hyperplane is informative with respect to classification certainty.

3. Evaluation

For the evaluation of our approach we manually labeled all possible links between the source and target Wikipedia documents. To score the outcome of the classifier, we first calculate the number of *True Positives* (TP: instances that get a positive label in both the reference and system output), *False positives* (FP: instances that are wrongly output by the system as being positive), *False Negatives* (FN: instances wrongly output by the system as being negative) and *True Negatives* (TN: instances that are considered to be negative by both the reference and the SVM classifier). These four classes were then used to calculate Precision, Recall and a weighted F-measure of Precision and Recall:

[7] http://code.google.com/apis/language/

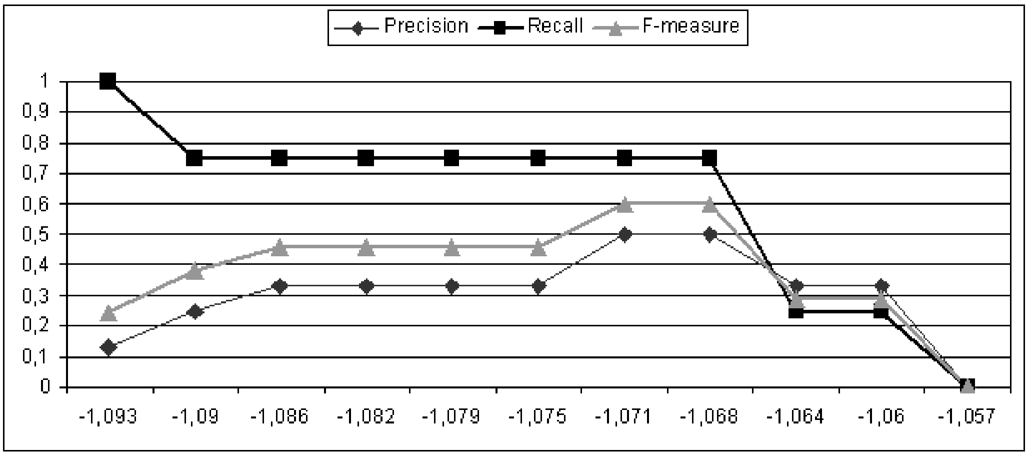

Figure 1: Dutch-French Precision, Recall and F-score for *muis* per threshold value for the output interval [-1.093,-1.057]

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

3.1. SVM Thresholds

SVMLIGHT outputs a floating point number for unseen instances: its sign designates the position, its absolute value the distance relative to the decision boundary. It is common practice to consider values above 0 as positive values, and values below 0 as negative values. However, the decision boundaries proposed by SVM classifiers are not necessarily optimal for all classification tasks. Previous research has shown that classification results can be strongly influenced by optimizing the decision boundary (Desmet and Hoste, 2012). Moving the decision boundary further away from the positive instances results in higher recall at the expense of precision, whereas moving it closer results in higher precision and lower recall figures. Changing the decision boundary can be done by defining a classification threshold other than 0.

Because we work with (1) very skewed data sets and (2) very sparse Bag-of-Words features to train the classifier, the SVM output is very irregular and not evenly distributed between [-1,1] as is the case for data sets with normal distributions between positive and negative examples. Therefore we decided to normalize the SVM output by considering the minimum and maximum SVM output values as the two boundaries of the classification range (which would be [-1, 1] in a default setup). Once these boundaries are identified, we divide the interval in slices of 10% and calculate Precision, Recall and the weighted F-score for these newly defined classification thresholds. To illustrate this, Figure 1 shows all Precision, Recall and F-measure results for the Dutch-French SVM output for *muis* (TF-IDF flavor of the feature vectors) with interval boundaries [-1.093,-1.057].

3.2. Results

Table 4 shows the experimental results for the feature vectors containing binary Bag-of-Words features, while Table 5 lists the results for the classifiers trained with weighted TF-IDF Bag-of-Words features. We have listed for each classifier the results when using the best SVM threshold, as well as the results when considering the middle threshold as the default threshold (and thus corresponding to 0 in a standard [-1,1] setup).

The experimental results lead to the following observations. First, the results confirm that moving the SVM decision boundary has a high impact on the classification results. The results obtained when using the optimal threshold clearly outperform the results when using the default threshold, although the performance gains are bigger for the classifier trained with binary BOW-features (optimal threshold results are 2.11 times better) than for the classifier trained with TF-IDF BOW-features. For the latter classifier, using the optimal threshold leads to an improvement of the F-score with 27% for Dutch-French. Furthermore, finding a good threshold appears to be obligatory, as for some words we do not obtain any True Positives at all when using the standard threshold (cfr. *operatie* in Dutch-Italian).

Second, the use of weighted Bag-of-Words features instead of binary features at the one hand leads to important performance gains (E.g. *muis* in Dutch-French and Dutch-Spanish), but on the other hand sometimes leads to performance drops (E.g. *stam*).

In addition, we performed a qualitative analysis in order to have a better insight in the performance of the system. The qualitative analysis shows that we indeed manage to find valid inter-language links and even succeed to detect missing links. The TF-IDF system for instance detects a link between the French *Souris* page and the Dutch *Muis (van de hand)*[8] page that is not present in Wikipedia. We detect even more important missing links for more frequent usages of the noun, such as the correspondences between the Dutch *Muis (animal)* and the Spanish and Italian corre-

[8] "ball of the thumb" sense of Muis

	Optimal Threshold			Middle Threshold		
	Prec	Rec	F-Score	Prec	Rec	F-Score
Dutch-French Results						
muis	0.33	0.25	0.29	0.07	0.25	0.11
graad	0.13	0.43	0.19	0.04	0.86	0.07
stam	0.21	0.27	0.24	0.07	0.64	0.13
operatie	0.13	0.67	0.21	0.08	0.33	0.13
Dutch-Italian Results						
muis	0.17	0.50	0.25	0.03	0.50	0.05
graad	0.11	0.57	0.19	0.06	0.57	0.11
stam	1.00	0.20	0.33	0.14	0.40	0.21
operatie	0.10	1.00	0.18	0.00	0.00	0.00
Dutch-Spanish Results						
muis	0.11	0.67	0.19	0.08	0.33	0.13
graad	0.05	0.75	0.10	0.03	0.25	0.06
stam	0.14	0.14	0.14	0.05	0.43	0.10
operatie	0.13	0.50	0.20	0.10	1.00	0.18

Table 4: Results for binary Bag-of-Words features

	Optimal Threshold			Middle Threshold		
	Prec	Rec	F-Score	Prec	Rec	F-Score
Dutch-French Results						
muis	0.50	0.75	0.60	0.33	0.75	0.46
graad	0.06	0.57	0.11	0.05	1.0	0.10
stam	0.10	0.36	0.15	0.06	0.73	0.11
operatie	0.25	0.33	0.29	0.10	0.67	0.17
Dutch-Italian Results						
muis	0.10	1.00	0.18	0.04	1.00	0.07
graad	0.06	1.00	0.12	0.06	0.71	0.12
stam	0.14	0.40	0.21	0.11	0.60	0.18
operatie	0.10	1.00	0.18	0.00	0.00	0.00
Dutch-Spanish Results						
muis	1.00	0.33	0.50	0.17	0.67	0.27
graad	0.04	1.00	0.08	0.00	0.00	0.00
stam	0.06	1.00	0.12	0.05	0.43	0.09
operatie	0.10	1.00	0.18	0.00	0.00	0.00

Table 5: Results for TF-IDF Bag-of-Words features

sponding pages, both for the SVM output with optimal and default thresholds.

Shallow error analysis revealed three reasons for incorrect (*False Positives*) or missing (*False Negatives*) links between document pairs. First, as we work on ambiguous nouns, there is sometimes overlap on the content words of two different meanings of the word (e.g. the Dutch word *hand* occurs in both the *mouse_"informatics"* and the *mouse_"ball of the thumb"* sense).

Second, the two corpora (Europarl and Wikipedia) that are used are very different in nature and vocabulary usage; Europarl is extracted from the proceedings of the European Parliament and contains transcriptions of political speeches, whereas Wikipedia typically contains encyclopedic knowledge and a more scientific vocabulary register. In order to improve the output of the system, we would ideally need to use a training corpus that better corresponds to the encyclopedic text genre of Wikipedia.

Third, we constructed the training corpus in an automatic way: aligned documents in Europarl (viz.Dutch documents and their aligned translations) are considered to be positive training instances, whereas all other document pairs are considered as negative training instances. It is possible, however, that some documents that are not aligned (i.e. that are no literal translations from each other) are also related to the same content and meaning of the ambiguous noun, and therefore might be labeled as positive training instances instead of negative ones.

As a final remark, we want to nuance the modest Precision figures. For the presented approach we aim to obtain higher Recall than Precision figures, as the final goal consists in providing a human editor with possible missing Wikipedia links that should be manually verified.

4. Conclusion

We presented a cross-lingual link discovery system that discovers missing Wikipedia inter-language links. Our system consists of two main modules: a *Greedy crawler* module that retrieves pages containing translations of the target word in the page titles, and a *Cross-lingual web page linker* that links corresponding pages in the input and target languages. We recasted the linking of two web pages as a classification problem, and used Bag-of-Words features (in a binary and TF-IDF weighted version) to find corresponding pages. The evaluation results show that we indeed succeed in detecting missing inter-language links between Wikipedia documents in Dutch and three target languages (viz. French, Italian and Spanish). Another important conclusion of the experiments is that optimizing the SVM decision boundary has a high impact on the classification results.

In future work, we would like to integrate the confidence scores of the SVM classifier in order to rank the possible links that are presented to the human editors. In addition, we will also compare the presented approach with an unsupervised clustering-based approach that is applied on concept-term matrices that are extracted from the considered Wikipedia pages.

5. References

S.F. Adafre and M. de Rijke. 2005. Discovering missing links in wikipedia. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 90–97, New York, USA.

R. Bunescu and M. Pasca. 2006. Using Encyclopedic Knowledge for Named entity Disambiguation. In *Proceedings of EACL 2006*, pages 47–52.

B. Desmet and V. Hoste. 2012. Combining Lexico-semantic Features for Emotion Classification in Suicide Notes. *Biomedical Informatics Insights*, 5:125–128.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

W.C. Huang, A. Trotman, and S. Geva. 2009. A Virtual Evaluation Forum for Cross Language Link Discovery. In *Proceedings of the SIGIR 2009 Workshop of the Future of IR Evaluation*, pages 19–20.

Thorsten Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*. Springer.

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

E. Lefever and V. Hoste. 2011. Examining the Validity of Cross-Lingual Word Sense Disambiguation. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*, Tokyo, Japan.

E. Lefever, V. Hoste, and M. De Cock. 2011. ParaSense or How to Use Parallel Corpora for Word Sense Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 317–322, Portland, Oregon, USA. Association for Computational Linguistics.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on information and knowledge management*, pages 233–242, New York, NY, USA.

V. Nastase. 2008. Topic-driven multi-document summarization with encyclopedic knowledge and activation spreading. In *Proceedings of EMNLP-08*, pages 763–772.

R. Navigli and S.P. Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

B. Raskutti and A. Kowalczyk. 2003. Extreme Rebalancing for SVMs: a case study. In *Proceedings of the Workshop on Learning from Imbalanced Datasets II*.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. In *Information Processing & Management*, volume 24(5), pages 513–523.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on new methods in Language Processing*, Manchester, UK.