Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4212–4220
Marseille, 20-25 June 2022

Medical Crossing: a Cross-lingual Evaluation of Clinical Entity Linking

Anton Alekseev[1,2], Zulfat Miftahutdinov[3], Elena Tutubalina[4,5], Artem Shelmanov[6], Vladimir Ivanov[7], Vladimir Kokh[8], Alexander Nesterov[8] Manvel Avetisian[8], Andrey Chertok[5,6], Sergey Nikolenko[1]

[1]Steklov Mathematical Institute at St. Petersburg, St. Petersburg, Russia [2]SPbSU, St. Petersburg, Russia
[3]Kazan Federal University, Kazan, Russia [4]HSE University, Moscow, Russia
[5]Sber AI, Moscow, Russia [6]AIRI, Moscow, Russia
[7]Innopolis University, Innopolis, Russia [8]Sber AI Lab, Moscow, Russia
{anton.m.alexeyev, zulfatmi, tutubalinaev, artemshelmanov, nomemm}@gmail.com
{kokh.v.n, ainesterov, avetisian.m.s, achertok}@sberbank.ru
sergey@logic.pdmi.ras.ru

Abstract

Medical data annotation requires highly qualified expertise. Despite the efforts devoted to medical entity linking in different languages, available data is very sparse in terms of both data volume and languages. In this work, we establish benchmarks for cross-lingual medical entity linking using clinical reports, clinical guidelines, and medical research papers. We present a test set filtering procedure designed to analyze the "hard cases" of entity linking approaching zero-shot cross-lingual transfer learning, evaluate state-of-the-art models, and draw several interesting conclusions based on our evaluation results.

Keywords: medical entity linking, embeddings, linking evaluation, cross-lingual methods, zero-shot learning

1. Introduction

Entity linking is the task of establishing correspondences between free-form text mentions and a formalized list of concepts (Shen et al., 2014; Sevgili et al., 2020). In this work, we consider *medical entity linking* – the task where entity mentions are mapped against a large set of medical concept names and their concept unique identifiers (CUIs). The biomedical domain is characterized by extensive dictionaries of concepts such as the Unified Medical Language System (UMLS) (Bodenreider, 2004), Medical Subject Headings (MeSH) (Coletti and Bleich, 2001), Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) (Spackman et al., 1997), or Medical Dictionary for Regulatory Activities (MedDRA) (Brown et al., 1999), and a high variation of mentions.

Early models for biomedical entity linking commonly used classification type losses (Rios and Kavuluru, 2018; Miftahutdinov and Tutubalina, 2019; Lou et al., 2020) that work well on narrow benchmarks but often lead to significant performance degradation on other domains and structurally different texts. Modern approaches usually employ similarity between *embeddings* (distributed representations) of words and concepts. From classical *tf-idf* and *word2vec* embeddings (Aronson, 2001; Ghiasvand and Kate, 2014; Van Mulligen et al., 2016; Leaman and Lu, 2016; Dermouche et al., 2016), entity linking systems have evolved to leverage vector representations constructed by deep neural models that take advantage of self-attention (Vaswani et al., 2017) and a BERT-like ranking architecture (Zhu et al., 2019; Sung et al., 2020; Tutubalina et al., 2020). We especially note the Biomedical Named Encoder (BNE) (Phan et al., 2019) and BioSyn (Sung et al., 2020) – a Transformer model based on BioBERT (Lee et al., 2020).

Along with the progress of embedding techniques and neural architectures, the reported performance of state-of-the-art entity linking models has been steadily increasing over the past years. However, their evaluation in many works remains limited. Oftentimes, models are evaluated in the *single-terminology* setting, on the same kind of data they were trained on, and in a very narrow domain devoted to a specific type of texts and/or a specific set of diseases (e.g. oncological) simply divided into training and test parts. Moreover, standard train/test splits often contain data leaks where the same terminology and even mentions of the same kind leak from the training set into the test set, significantly improving the scores and restricting fair evaluation of transfer capabilities to other domains. Tutubalina et al. (2020) show that this effect leads to a significant positive bias in reported quality metrics and that such data leaks do exist in biomedical datasets of English scientific abstracts widely used for entity linking evaluation.

Lately, entity linking has started to shift towards the *zero-shot* setting, where the test set contains only novel concepts that have not been seen in the training data (Logeswaran et al., 2019; Basaldella et al., 2020; Mohan et al., 2021; Sevgili et al., 2020). This setting is harder and can be considered more "fair" since it mitigates many trivial linking cases. In this work, unlike commonly used *single-terminology* evaluation, where all concept names and CUIs from a target dictionary are seen during training, we consider a *cross-terminology* setting – a sophisticated version of *zero-shot*: test sets contain novel concepts from a target terminology, while another terminology is used during training.

A recent systematic literature survey (Kersloot et al., 2020) reviews the current state of the development and evaluation of NLP algorithms for mapping medical text fragments onto ontology concepts. The authors study 77 works, and only 17 (22%) of them perform the evaluation on non-English datasets, including Italian (Combi et al., 2018), Portuguese (Duarte et al., 2018), Japanese (Usui et al., 2018), and Korean (Kang et al., 2008). Although those datasets do not always contain entity linking annotations, there is an imbalance of English/non-English data. Moreover, prior art with cross-terminology evaluation has been restricted to the single-language setting. In this work, we make another step towards the fair evaluation of medical entity linking models across languages. We unite these two directions, providing both cross-terminology and cross-lingual evaluation on real-life biomedical and clinical texts.

We test the transfer capabilities of recently proposed models for medical entity linking across languages, taking care to avoid leaks from training to test parts of the datasets used. We seek to answer the following research questions:

RQ1: Do test sets of current benchmarks in English, Spanish, French, German, and Dutch lead to an overestimation of performance?

RQ2: What is the fair evaluation strategy?

RQ3: What is the potential of a model trained on a corpus in English to generalize for the zero-shot clinical entity linking in other languages?

RQ4: What types of word representations can be used for cross-lingual clinical entity linking (state-of-the-art contextual word representations, sparse representations)?

We show that filtering the test sets to avoid leaks proves to be crucial for a fair evaluation and provides new interesting and sometimes unexpected conclusions: sparse baselines consistently outperform BERT-based models, domain knowledge is very important for the quality, and fine-tuning on medical datasets can significantly improve the results, an effect that is not noticeable in common benchmarks without filtering.

2. Data

We construct a full-scale multilingual evaluation benchmark from several real-life clinical and biomedical datasets. Table 1 summarizes basic statistics of these datasets: number of concepts, number and the average length of entity mentions, percentage of mentions with numerals. Examples of dataset instances are presented in Table 2.

2.1. CodiEsp

The *CodiEsp* dataset was presented at Clinical Case Coding in Spanish Shared Task at the CLEF 2020 evaluation forum (Miranda-Escalada et al., 2020b). It contains structured information (clinical records) with entities mapped against the ICD-10 vocabulary (CodeBooks, 2016); we use the CodiEsp Diagnosis (CodiEsp-D) subset and the dictionary provided in *CodiEsp*.

2.2. Cantemist

Cantemist (CANcer TExt MIning Shared Task on IberLEF 2020 (Miranda-Escalada et al., 2020a)) is a manually annotated text corpus of tumor morphology mentions in Spanish mapped to the latest Spanish version of the oncological ontology, which is a part of ICD-O (World Health Organization, 2013); we use the dictionary from (López-Úbeda et al., 2020).

2.3. MCN

MCN (Medical Concept Normalization) (Luo et al., 2019) is a large-scale manually annotated corpus in English for clinical concept normalization produced from a corpus released for the 4th i2b2/VA shared task (Uzuner et al., 2011) with a dictionary of concepts from SNOMED-CT extracted from the UMLS 2020 AA release.

2.4. Mantra

Mantra GSC (Kors et al., 2015) is a collection of biomedical text units such as drug labels and patent claims manually cross-labeled by several annotators in five different languages: English, French, German, Spanish, and Dutch. The Mantra terminology is a subset of UMLS with concepts from MeSH, SNOMED-CT, and MedDRA extracted from the UMLS 2020 AA release; we use DISO entities (UMLS semantic group "Disorders" (Bodenreider and McCray, 2003)).

2.5. Other Datasets

Other available clinical datasets do not suit our needs. The German clinical guidelines dataset (Borchert et al., 2020) does not have concept-level annotations. English, Spanish, and Portuguese texts in Multi-NEL (Ruas et al., 2020) are synthetic. The Portuguese clinical notes dataset (Peters et al., 2020), the Japanese dataset of patient complaints (Usui et al., 2018), the Korean clinical dataset (Kang et al., 2008), and the Italian drug reaction corpus (Combi et al., 2018) are not publicly available yet. The dataset of death certificates in Portuguese does not contain annotated entities and is not publicly available (Duarte et al., 2018).

An important recent work presented the XL-BEL cross-lingual biomedical entity linking task (Liu et al., 2021) that allowed to test domain transfer across languages. However, XL-BEL does not allow for cross-terminology transfer evaluation and basically represents *WikiMed* (Vashishth et al., 2020) aligned across

Dataset	Lang	# in full corpus	Avg. len in chars	% with numerals	Split		Filtering			
					Train	Test	Train set		Dictionary	
							Filt.	$\text{Filt}_{0.2}$	Filt.	$\text{Filt}_{0.2}$
Entity mentions										
CANTEMIST	es	10031	18.73	6.92	6396	3635	998	711	3268	3040
CodiEsp-D	es	10874	15.84	1.05	7209	3665	1386	1167	3449	3347
MCN	en	13609	12.36	1.54	6684	6925	3204	2819	3386	2304
Mantra	de	201	17.62	0.50	-	201	-	-	107	62
	en	452	16.42	1.11	-	452	-	-	126	66
	es	166	19.67	2.41	-	166	-	-	65	38
	fr	222	17.64	0.45	-	222	-	-	99	50
	nl	127	16.06	0.00	-	127	-	-	65	44
Concepts										
CANTEMIST	es	657	-	-	493	386	332	279	364	321
CodiEsp-D	es	2206	-	-	1767	1143	841	750	1142	1050
MCN	en	3792	-	-	2331	2579	2000	1834	1631	1195
Mantra	de	169	-	-	-	169	-	-	97	53
	en	373	-	-	-	373	-	-	119	61
	es	147	-	-	-	147	-	-	69	35
	fr	185	-	-	-	185	-	-	83	39
	nl	117	-	-	-	117	-	-	62	42

Table 1: Statistics of the datasets in English (en), Spanish (es), French (fr), German (de), and Dutch (nl).

ten different languages via *Wikipedia*, so the critique above fully applies to XL-BEL as well. We note an important difference between datasets such as *WikiMed* (Vashishth et al., 2020) and medical texts such as clinical health records or scientific abstracts. The usage of medical terms is very different between *Wikipedia* and other texts, so entity linking results may not transfer well. In this work, we use a disease-centric approach to data collection, with a broad collection of datasets with real medical texts.

2.6. Filtering Strategies

We present a novel test set *filtering* strategy to avoid train/test leaks and provide a fair and more challenging comparison in the cross-terminology setting. We construct a reference set of terms from concept names in an entity dictionary (thesaurus) and filter out from the test set all instances, in which mention surface forms match any term in the reference set (*filtering by a dictionary*). We also perform the evaluation in a less challenging setting suggested by Tutubalina et al. (2020) where the reference set for filtering is constructed from the entity mentions in the training dataset (*filtering by a training set*). For a reference set of terms/entities, we provide the following evaluation types:

- *Full*: compute metrics on the test set as provided in the dataset itself;
- *Filtered*: remove from the test set all entities that are already present in the reference set (exact match, e.g., we remove all instances of "depression" from the test set if it is already present in the reference set);
- *Filtered*$_{0.2}$: remove from the test set all entities where the character-based Levenshtein distance to the nearest neighbor in the reference set is under 0.2 (e.g., we remove "depressed" if "depression" occurs in the reference set). This complicates the task even further since a model cannot rely on word similarity and have to use more sophisticated contextual features. The bigger the threshold the harder the evaluation setting.

Table 1 shows how many concepts and entity mentions remain in the test sets of each of the datasets after the corresponding filtering method is applied. Note that filtering significantly reduces the number of entity mentions in test sets across all datasets, and the difference is especially striking for training set filtering. This indicates a large number of train set leaks that we discussed in Section 1.

3. Models for Medical Entity Linking

For entity linking, we use a ranking model based on embeddings of a mention and a possible concept. Each entity mention and a concept name is passed first through a model that produces their embeddings and then through an average pooling layer that yields a fixed-sized vector. The inference task is then reduced to finding the closest concept name representation to entity mention representation in a common embedding space, where the Euclidean distance can be used as the metric. Nearest concept names are chosen as top-k concepts for entities.

3.1. Entity and Concept Representations

We compare the following mention/entity vector representations:

- *Tf-idf*: standard sparse *tf-idf* representations constructed from character-level unigrams and bigrams;

Dataset	Lang	Name	CUI	Mention
CANTEMIST	es	“Neoplasia maligna”	8000/3	malignidad
		...malignos o de **malignidad** intermedia...		
		“Neoplasia metastásica”	8000/6	metastásico
		...compromiso **metastásico**, y tras presentarse...		
CodiEsp-D	es	“otros trastornos especificados de músculo”	M62.89	hipertrofia del psoas
		“adenomegalia localizada”	R59.0	Adenopatías inguinales
MCN	en	“Gastritis”, “Gastric catarrh”, etc.	C0017152	gastritis
		...was negative for **gastritis** , stricture or ulcer...		
		“Empirical therapy (procedure)”	C1299597	empiric treatment
		...was started on **empiric treatment**...		
Mantra (DISO)	de	“Arthralgie”, “Gelenkschmerz”, etc.	C0003862	arthralgien
		...Übelkeit, **Arthralgien**, niedrigem Blutdruck...		
		“Lumbalgie”, “Unterer Rueckenschmerz”, etc.	C0024031	kreuzschmerzen
		...und mittelstarken **Kreuzschmerzen** kommen...		
	en	“Nausea (disorder)”, “Feeling queasy”, etc.	C0027497	nausea
		“Arthralgia”, “Pain in joint”, etc.	C0003862	arthralgia
		...reactions, **nausea**, **arthralgia**, low blood pressure...		
	es	“Inflamación pulmonar”, “Neumonía”, etc.	C0032285	neumonía
		...**Neumonía**⋆, infección de vías respiratorias...		
		“Infección de los senos”, “Sinusitis”, etc.	C0037199	sinusitis
		...respiratorias altas, **sinusitis**, candidiasis oral...		
	fr	“Anoréxique”, “Anorexie”, etc.	C0003123	anorexie
		...incluent fièvre, **anorexie** (perte d’ appétit)...		
		“Irritabilité”, “Humeur irritable”, etc.	C0022107	irritabilité
		...vomissements, diarrhée, **irritabilité**, somnolence...		
	nl	“blaasneoplasma”, “neoplasma blaas”, etc.	C0005695	blaastumoren
		...classificatie van **blaastumoren** en de behandeling...		
		“weefsel infiltratie”	C0332448	infiltrerende
		...de oppervlakkig **infiltrerende** tumoren...		

Table 2: Data samples from test sets (with fragments of original source texts where available). Each contains a mention (e.g. “sinusitis”) and a concept ID (e.g. “C0037199”). Note that identifiers come from different sets. “Names” are taken from: “valid_codes.txt” (a list of codes and names provided by the competition organizers) for *Cantemist*, “codiesp_codes” (a list of valid CIE10 codes provided by the CLEF2020 eHealth track organizers as a dictionary for the corresponding task) for *CodiEsp*, SNOMEDCT_US part of UMLS for *MCN* and *Mantra-En*, the rest are taken from MedDRA in German, Spanish, French, and Dutch, respectively.

- *BERT*: multilingual BERT embeddings with no fine-tuning (Devlin et al., 2019); this is a cross-lingual baseline that has not been trained on biomedical texts;
- *BETO*: Spanish BERT embeddings (Cañete et al., 2020);
- *BioBERT-esp*: BioBERT embeddings fine-tuned over Spanish clinical data (Villena, 2021) (we test BioBERT-esp and BETO on Spanish datasets);
- *SapBERT*: a BERT-based metric learning framework that generates hard triplets based on the UMLS for large-scale pre-training (Liu et al., 2021a) and also allows for a cross-lingual variant (Liu et al., 2021b) trained on XL-BEL (Liu et al., 2021).

3.2. Fine-tuning

To fine-tune SapBERT models, we use synonym marginalization and iterative candidate retrieval as suggested in a recent state-of-the-art model *BioSyn* (Sung et al., 2020). We compare the following versions:

- *SapBERT+target* with fine-tuning on the target train set;
- *SapBERT+mcn* with fine-tuning on the MCN English train set;
- *SapBERT+mcn-fz4* and *SapBERT+mcn-fz10* on the MCN English training set with freezing the first four and ten layers, respectively.

4. Experiments

4.1. Experimental Setup

For monolingual evaluation, we leverage the train / test splits provided with each corpus. As shown in Table 1, only CANTEMIST, CodiEsp, and MCN have a train/test split in our study: Mantra subsets are too small for fine-tuning. For cross-lingual evaluation, we train models on the MCN English train set with a source dictionary and evaluate on the test sets of each other corpora (i.e., the *target*). Specifically, ranking models retrieve the nearest concept name in a target dictionary for a given mention representation at the inference time. We note that cross-lingual evaluation pro-

Dataset	Model	Full		Filtered		Filtered$_{0.2}$	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
CodiEsp Diagnostico	Tf-idf	20.55%	39.24%	14.21%	25.76%	13.62%	24.51%
	BERT	10.45%	15.58%	6.49%	9.88%	6.51%	9.68%
	BETO	9.47%	15.09%	5.92%	10.03%	5.83%	10.03%
	BioBERT-esp	10.07%	14.38%	6.78%	11.98%	7.11%	12.34%
	SapBERT	47.83%	63.66%	32.61%	46.10%	31.62%	45.33%
	SapBERT+target	67.18%	76.23%	47.62%	61.26%	45.42%	58.53%
	SapBERT+mcn	48.27%	64.07%	33.04%	47.69%	31.96%	46.19%
	SapBERT+mcn-fz4	48.32%	63.68%	33.48%	47.40%	32.56%	45.76%
	SapBERT+mcn-fz10	49.14%	64.31%	33.26%	47.76%	31.62%	45.67%
MCN	Tf-idf	59.00%	65.91%	52.12%	62.77%	51.15%	61.58%
	BERT	48.61%	52.16%	36.64%	41.29%	36.64%	41.15%
	SapBERT	66.28%	74.55%	62.84%	71.99%	59.95%	69.03%
	SapBERT+target	69.36%	80.90%	66.94%	74.42%	63.64%	73.79%
CANTEMIST	Tf-idf	27.02%	47.92%	20.24%	31.76%	20.25%	32.07%
	BERT	25.50%	34.69%	8.72%	13.43%	8.72%	13.50%
	BETO	13.43%	19.17%	9.82%	14.13%	10.13%	14.77%
	BioBERT-esp	15.24%	23.41%	11.72%	18.94%	11.81%	19.13%
	SapBERT	57.47%	65.23%	28.06%	36.47%	28.41%	36.99%
	SapBERT+target	79.45%	87.76%	53.31%	68.54%	51.48%	66.10%
	SapBERT+mcn	61.29%	67.02%	29.06%	39.98%	29.54%	40.51%
	SapBERT+mcn-fz4	61.60%	66.63%	29.66%	39.28%	30.10%	40.23%
	SapBERT+mcn-fz10	57.47%	65.45%	28.06%	37.27%	28.55%	37.41%

Table 3: Results of the evaluation with filtering by a training set.

vides a challenging setup for the standard supervised models, especially for linking of mentions in another language not encountered during training.

We evaluate the models in the information retrieval scenario, where the goal is to find top-k concepts for every entity mention in a dictionary of concept names and their identifiers. Following previous works on entity linking (Suominen et al., 2013; Pradhan et al., 2014; Wright et al., 2019; Phan et al., 2019; Sung et al., 2020; Tutubalina et al., 2020), we use the top-k accuracy as the evaluation metric: $\mathrm{Acc@k} = 1$ if the correct UMLS concept unique identifier is retrieved at the rank $\leq k$, otherwise $\mathrm{Acc@k} = 0$.

For evaluation of methods that perform ranking without fine-tuning, we leverage publicly available implementation from (Tutubalina et al., 2020)[1] and the following pre-trained models available in the Hugging Face (Wolf et al., 2020) repository:

- BERT-multilingual (Devlin et al., 2019): `bert-base-multilingual-cased`;
- BETO (Cañete et al., 2020): `dccuchile/bert-base-spanish-wwm-uncased`;
- BioBERT-esp (Villena, 2021) `fvillena/bio-bert-base-spanish-wwm-uncased`.

The implementation of the core SapBERT is based on the publicly available repository (Sung et al., 2020)[2]. The modifications are taken from the public BioSyn repository[3]. We fine-tune various SapBERT models (Liu et al., 2021b) starting from the pre-trained checkpoint `SapBERT-UMLS-2020AB-all-lang-from-XLMR`, which was constructed by the authors from cross-lingual RoBERTa (Conneau et al., 2019), `xlm-roberta-base`. The pre-training hyperparameters for SapBERT can be found in the original work. We performed the fine-tuning with the following hyperparameters: the number of top candidates k is 20, the mini-batch size is 16, the learning rate is 1e-5, the dense ratio for candidate retrieval is 0.5.

4.2. Results

Table 3 shows the Acc@1 and Acc@5 metrics for datasets with the training set used as the reference set for filtering, while Table 4 shows these variations with the entity dictionary used as the reference set for filtering. Table 3 does not contain the *Mantra* dataset because it is too small to reasonably use for fine-tuning.

The results of our evaluation suggest several important and interesting conclusions.

First, Tables 3 and 4 show a significant difference between evaluation strategies: on full test sets, there is virtually no difference between SapBERT variations, but on filtered datasets, fine-tuning on MCN or the target dataset brings a significant increase in accuracy. For weaker baselines, the filtering effect can be drastic. For example, note how the BERT-based model in Table 4 dropped from 48% top-1 accuracy to 12.5% and 6.2% on the MCN dataset after dictionary-based filtering. This indicates that the most successful matches

[1] `https://github.com/insilicomedicine/Fair-Evaluation-BERT`

[2] `https://github.com/cambridgeltl/sapbert`

[3] `https://github.com/dmis-lab/BioSyn`

Dataset	Model	Full		Filtered		Filtered$_{0.2}$	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
CodiEsp Diagnostico	Tf-idf	20.55%	39.24%	15.63%	35.49%	15.45%	35.28%
	BERT	10.45%	15.58%	4.90%	10.35%	4.75%	10.18%
	SapBERT	47.83%	63.66%	44.62%	61.44%	44.55%	61.14%
	SapBERT+mcn	48.27%	64.07%	45.09%	61.87%	44.19%	60.98%
	SapBERT+mcn-fz4	48.32%	63.68%	45.14%	61.47%	44.25%	60.56%
	SapBERT+mcn-fz10	49.14%	64.31%	46.01%	62.13%	38.54%	50.95%
MCN	Tf-idf	59.00%	65.91%	33.82%	45.87%	24.61%	36.55%
	BERT	48.61%	52.16%	12.55%	19.46%	6.21%	10.98%
	SapBERT	66.28%	74.55%	47.50%	59.08%	38.54%	50.80%
	SapBERT+target	69.36%	80.90%	54.99%	67.13%	46.14%	58.16%
CANTEMIST	Tf-idf	27.02%	47.92%	18.85%	42.07%	16.57%	28.01%
	BERT	25.50%	34.69%	17.17%	27.36%	16.48%	26.55%
	SapBERT	57.47%	65.23%	52.72%	61.32%	51.12%	59.64%
	SapBERT+mcn	61.29%	67.02%	56.98%	63.31%	55.86%	61.61%
	SapBERT+mcn-fz4	61.6%	66.36%	57.31%	62.88%	56.22%	61.05%
	SapBERT+mcn-fz10	57.47%	65.45%	52.72%	61.57%	51.12%	59.64%
Mantra (German)	Tf-idf	73.63%	79.10%	50.47%	60.75%	29.03%	45.16%
	BERT	59.20%	63.68%	23.36%	31.78%	8.07%	16.13%
	SapBERT	87.56%	95.52%	76.64%	91.59%	64.52%	88.71%
	SapBERT+mcn	88.06%	95.52%	80.30%	89.39%	67.74%	87.10%
	SapBERT+mcn-fz4	89.55%	95.02%	80.37%	90.65%	72.58%	87.10%
	SapBERT+mcn-fz10	88.06%	95.52%	77.57%	91.59%	66.13%	88.71%
Mantra (English)	Tf-idf	86.06%	92.04%	51.59%	73.02%	43.94%	62.12%
	BERT	78.54%	84.29%	24.60%	45.24%	16.67%	37.88%
	SapBERT	93.81%	96.90%	79.37%	90.48%	75.76%	90.91%
	SapBERT+mcn	94.03%	96.90%	80.16%	90.48%	80.30%	89.39%
	SapBERT+mcn-fz4	94.25%	97.12%	80.95%	91.27%	80.16%	90.48%
	SapBERT+mcn-fz10	94.25%	96.90%	80.95%	90.48%	80.30%	90.91%
Mantra (Spanish)	Tf-idf	71.69%	80.72%	45.45%	62.34%	26.32%	44.74%
	BERT	62.65%	69.28%	25.97%	38.96%	10.53%	15.79%
	SapBERT	83.73%	90.36%	71.43%	83.12%	47.37%	68.42%
	SapBERT+mcn	84.34%	90.96%	72.73%	84.42%	50.00%	71.05%
	SapBERT+mcn-fz4	85.54%	92.17%	75.32%	87.01%	52.63%	76.32%
	SapBERT+mcn-fz10	84.34%	92.77%	72.73%	87.01%	47.37%	76.32%
Mantra (French)	Tf-idf	77.03%	80.63%	50.51%	57.58%	30.00%	38.00%
	BERT	65.32%	71.62%	24.24%	37.37%	2.00%	12.00%
	SapBERT	82.43%	93.24%	62.63%	84.85%	46.00%	76.00%
	SapBERT+mcn	83.33%	95.50%	64.65%	89.90%	54.00%	84.00%
	SapBERT+mcn-fz4	84.23%	94.14%	66.67%	86.87%	54.00%	80.00%
	SapBERT+mcn-fz10	82.88%	93.69%	63.64%	85.86%	48.00%	78.00%
Mantra (Dutch)	Tf-idf	73.23%	77.95%	53.85%	61.54%	43.18%	50.00%
	BERT	55.12%	58.27%	18.46%	24.62%	13.64%	20.45%
	SapBERT	84.25%	87.40%	73.85%	80.00%	63.64%	72.73%
	SapBERT+mcn	85.83%	87.40%	78.46%	80.00%	70.45%	72.73%
	SapBERT+mcn-fz4	85.83%	87.40%	78.46%	80.00%	70.45%	72.73%
	SapBERT+mcn-fz10	84.25%	87.40%	75.38%	80.00%	65.91%	72.73%

Table 4: Results of the evaluation with filtering by a dictionary.

of these models come from training set leaks and very simple cases of entity linking (surface forms). A fair comparison requires filtering procedures such as the ones we suggest in this paper.

Another result is that fine-tuning on additional medical data is generally beneficial; e.g., we have found that SapBERT fine-tuned on English clinical notes outperforms basic SapBERT consistently across all datasets in our study. However, a separate experimental evaluation is required to find the best parameters for this process: which layers to freeze during fine-tuning, how many epochs of training to conduct, etc. Interestingly, fine-tuning SapBERT improves results only after one epoch (we show these in the tables), and then the quality begins to drop, probably signifying overfitting. We also note that fine-tuning on the target dataset instead of English MCN as expected helps to substantially improve the quality.

Finally, the weaker baselines also provide new insights. The sparse *tf-idf* baseline consistently outperforms BERT-based ranking. Many recent works forgo sparse baselines entirely, but our results suggest that it may be premature. Both multilingual and Spanish BERT consistently perform much worse than all competitors, showing that biomedical domain knowledge is crucial for solving this task.

5. Conclusion

We have presented the first cross-lingual benchmark for clinical entity linking in English, Spanish, French, German, and Dutch. We perform an extensive evaluation of BERT-based models with state-of-the-art biomedical representations in two setups: with official train/test splits and with filtered test sets. Our filtering strategy keeps only entity mentions, which are dissimilar to entries from the reference set. As the reference set, we adopt a training set or a target entity dictionary. Our evaluation shows the great divergence in performance between official and proposed test sets for all languages and models, answering positively to the **RQ1** and supporting the claim that fair evaluation requires the proposed dataset filtering (the answer to the **RQ2**). Our experiments with SapBERT show that cross-lingual training on the English MCN corpus substantially helps to improve the performance on clinical datasets in other languages, which answers the **RQ3**. Finally, answering the **RQ4**, we note that general-purpose models without domain knowledge and fine-tuning are almost useless for the considered task, falling behind even the simplistic tf-idf baseline. Our fair evaluation shows that clinical entity linking requires pretraining at least on the related biomedical corpora. The constructed benchmark for cross-lingual clinical entity linking is available at `https://github.com/AIRI-Institute/medical_crossing`.

Our study opens up new venues for further work. First, we plan to extend this evaluation to more languages, more corpora, and other types of entities (not only diseases but, e.g., medical procedures or drugs). Second, SapBERT receives a significant boost in the performance by using synonymous relations, but in fact, the concepts form a tree-like hierarchy, and taking it into account may improve the results further. Third, since our method of evaluation moves towards zero-shot territory, we plan to apply other recently developed approaches in zero-shot learning to the entity linking problem.

6. Acknowledgements

We would also like to thank the anonymous reviewers for the comments and suggestions, which helped us improve the manuscript. The work is supported by the Russian Science Foundation [grant number 18-11-00284].

7. Bibliographical References

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Basaldella, M., Liu, F., Shareghi, E., and Collier, N. (2020). Cometa: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Combi, C., Zorzi, M., Pozzani, G., Moretti, U., and Arzenton, E. (2018). From narrative descriptions to meddra: automagically encoding adverse drug reactions. *Journal of Biomedical Informatics*, 84:184–199.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Dermouche, M., Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., and Taright, N. (2016). ECSTRA-INSERM@ CLEF eHealth2016-task 2: ICD10 code extraction from death certificates. CLEF.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Duarte, F., Martins, B., Pinto, C. S., and Silva, M. J. (2018). Deep neural models for icd-10 coding of death certificates and autopsy reports in free-text. *Journal of biomedical informatics*, 80:64–77.

Ghiasvand, O. and Kate, R. J. (2014). Uwm: Disorder mention extraction from clinical text using crfs and normalization using learned edit distance patterns. In *SemEval@ COLING*, pages 828–832.

Kang, B.-Y., Kim, D.-W., and Kim, H.-G. (2008). Two-phase chief complaint mapping to the umls metathesaurus in korean electronic medical records. *IEEE Transactions on Information Technology in Biomedicine*, 13(1):78–86.

Kersloot, M. G., van Putten, F. J., Abu-Hanna, A., Cornet, R., and Arts, D. L. (2020). Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. *Journal of biomedical semantics*, 11(1):1–21.

Leaman, R. and Lu, Z. (2016). Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. (2021a). Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, June.

Liu, F., Vulić, I., Korhonen, A., and Collier, N. (2021b). Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574.

Logeswaran, L., Chang, M.-W., Lee, K., Toutanova, K., Devlin, J., and Lee, H. (2019). Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460.

Lou, Y., Qian, T., Li, F., Zhou, J., Ji, D., and Cheng, M. (2020). Investigating of disease name normalization using neural network and pre-training. *IEEE Access*, 8:85729–85739.

Miftahutdinov, Z. and Tutubalina, E. (2019). Deep neural models for medical concept normalization in user-generated texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 393–399.

Mohan, S., Angell, R., Monath, N., and McCallum, A. (2021). Low resource recognition and linking of biomedical concepts from a large ontology. *arXiv preprint arXiv:2101.10587*.

Phan, M. C., Sun, A., and Tay, Y. (2019). Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285.

Pradhan, S., Elhadad, N., Chapman, W. W., Manandhar, S., and Savova, G. (2014). Semeval-2014 task 7: Analysis of clinical text. In *SemEval@ COLING*, pages 54–62.

Rios, A. and Kavuluru, R. (2018). Emr coding with semi–parametric multi–head matching networks. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2081. NIH Public Access.

Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., and Biemann, C. (2020). Neural entity linking: A survey of models based on deep learning. *arXiv preprint arXiv:2006.00575*.

Shen, W., Wang, J., and Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Sung, M., Jeon, H., Lee, J., and Kang, J. (2020). Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J., et al. (2013). Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Tutubalina, E., Kadurin, A., and Miftahutdinov, Z. (2020). Fair evaluation in concept normalization: a large-scale comparative analysis for bert-based models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6710–6716.

Usui, M., Aramaki, E., Iwao, T., Wakamiya, S., Sakamoto, T., and Mochizuki, M. (2018). Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: Natural language processing analysis in japanese. *JMIR medical informatics*, 6(3):e11021.

Van Mulligen, E., Afzal, Z., Akhondi, S. A., Vo, D., and Kors, J. A. (2016). Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts. CLEF.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Villena, F. (2021). Spanish biobert embeddings.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wright, D., Katsis, Y., Mehta, R., and Hsu, C.-N. (2019). Normco: Deep disease normalization for biomedical knowledge base construction. In *Automated Knowledge Base Construction*.

Zhu, M., Celikkaya, B., Bhatia, P., and Reddy, C. K. (2019). Latte: Latent type modeling for biomedical entity linking. *arXiv preprint arXiv:1911.09787*.

8. Language Resource References

Bodenreider, Olivier and McCray, Alexa T. (2003). *Exploring semantic groups through visual approaches*. Elsevier.

Bodenreider, Olivier. (2004). *The unified medical language system (UMLS): integrating biomedical terminology*. Oxford University Press.

Borchert, Florian and Lohr, Christina and Modersohn, Luise and Langer, Thomas and Follmann, Markus and Sachs, Jan Philipp and Hahn, Udo and Schapranow, Matthieu-P. (2020). *GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines*.

Brown, Elliot G and Wood, Louise and Wood, Sue. (1999). *The medical dictionary for regulatory activities (MedDRA)*. Springer.

CodeBooks, Medical. (2016). *ICD-10-CM Complete Code Set 2016*. Medical Code Books.

Coletti, Margaret H and Bleich, Howard L. (2001). *Medical subject headings used to search the biomedical literature*. BMJ Group BMA House, Tavistock Square, London, WC1H 9JR.

Jan A. Kors and S. Clematide and Saber Ahmad Akhondi and Erik M. van Mulligen and Dietrich Rebholz-Schuhmann. (2015). *A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC*.

Fangyu Liu and Ivan Vulic and Anna Korhonen and Nigel Collier. (2021). *Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking*.

López-Úbeda, Pilar and Díaz-Galiano, Manuel Carlos and Martín-Valdivia, María Teresa and López, Luis Alfonso Ureña. (2020). *Extracting Neoplasms Morphology Mentions in Spanish Clinical Cases through Word Embeddings.*

Yen-Fu Luo and Weiyi Sun and Anna Rumshisky. (2019). *MCN: A comprehensive corpus for medical concept normalization*.

Miranda-Escalada, Antonio and Farré, Eulàlia and Krallinger, Martin. (2020a). *Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results.*

Miranda-Escalada, Antonio and Gonzalez-Agirre, Aitor and Armengol-Estapé, Jordi and Krallinger, Martin. (2020b). *Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020*.

Peters, Ana Carolina and da Silva, Adalniza Moura Pucca and Gebeluca, Caroline P and Gumiel, Yohan Bonescki and Cintho, Lilian Mie Mukai and Carvalho, Deborah Ribeiro and Hasan, Sadid A and Moro, Claudia Maria Cabral and others. (2020). *SemClinBr–a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks*.

Pedro Ruas and André Lamúrias and Francisco M. Couto. (2020). *Towards a Multilingual Corpus for Named Entity Linking Evaluation in the Clinical Domain*. CEUR-WS.org.

Spackman, Kent A and Campbell, Keith E and Côté, Roger A. (1997). *SNOMED RT: a reference terminology for health care.*

Uzuner, Ozlem and South, Brett and Shen, Shuying and DuVall, Scott. (2011). *2010 i2B2/VA challenge on concepts, assertions, and relations in clinical text*.

Vashishth, Shikhar and Joshi, Rishabh and Newman-Griffis, Denis and Dutt, Ritam and Rose, Carolyn. (2020). *MedType: Improving Medical Entity Linking with Semantic Type Prediction*.

World Health Organization. (2013). *International classification of diseases for oncology (ICD-O)*. 3rd edition, 1st revision.