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 SPb, [2]SPbSU, [3]KFU, [4]HSE,
[5]Sber AI, [6]AIRI, [7]Innopolis University, [8]Sber AI Lab

Motivation

The reported performance of medical entity linking (EL) systems has been steadily improving, but their evaluation in many works is limited to narrow domains / single languages and corrupted by data leaks. We present:

1. A benchmark for x-lingual medical EL using clinical reports, clinical guidelines, and medical research papers
2. A test set filtering procedure designed to analyze the "hard cases" of EL approaching **0**-shot x-lingual transfer learning
3. SoTA EL model evaluation
4. Interesting conclusions drawn from evaluation on our benchmark

RQ1: Do current benchmarks in EN, ES, FR, DE, and NL lead to an overestimation of performance?

RQ2: What is the fair evaluation strategy for clinical entity linking (EL)?

RQ3: What is the potential of a model trained on English to generalize for **0**-shot clinical EL in other languages?

RQ4: What types of word representations can be used for cross-lingual clinical EL (SoTA contextual word vectors, sparse representations)?

Datasets

Medical datasets originating from real clinical records (CANTEMIST, CodiEsp, MCN) and drug labels, patent claims (Mantra GSC), etc. Dataset contains:

- mentions of entities linkable to standard ontologies,
- corresponding entities IDs, i.e. CUIs,
- [optional] original texts/contexts.

Languages: English, Spanish, French, German, and Dutch

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...		

Proposed Evaluation Settings

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 (a) concept names in an entity dictionary (thesaurus) or (b) from the entity mentions in the training dataset (less challenging setup).

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 already present in the reference set (**exact match**, e.g., removing instances of *depression* from the test set if already present in the reference set);
- **Filtered$_{0.2}$**: remove from the test set all entities where the normalized character-based Levenshtein distance to the nearest neighbor in the reference set is under 0.2 (e.g. removing *depressed* if *depression* occurs in the reference set). This makes the task more challenging since a model **cannot rely on word similarity** and have to use more sophisticated contextual features.

Filtering Results

Dataset	Lang	# in full corpus	Avg. len in chars	% with numerals	Split Train	Split Test	Filtering Train set Filt.	Filtering Train set Filt$_{0.2}$	Filtering Dictionary Filt.	Filtering Dictionary 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

Evaluation Results

- Evaluation on the official test sets and test sets filtered by a training set (removed all mentions from the training set)

Dataset	Model	Full Acc@1	Full Acc@5	Filtered Acc@1	Filtered Acc@5	Filtered$_{0.2}$ Acc@1	Filtered$_{0.2}$ 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%

- Evaluation on the official test sets and test sets filtered by an entity dictionary (more challenging)

Dataset	Model	Full Acc@1	Full Acc@5	Filtered Acc@1	Filtered Acc@5	Filtered$_{0.2}$ Acc@1	Filtered$_{0.2}$ 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%

- Great divergence in performance: official vs filtered test sets for all languages and models (positive answer to **RQ1** + claim that "fair" evaluation requires the proposed filtering is supported (**RQ2**)
- SapBERT experiments: cross-lingual training on the English MCN corpus improves the performance in other languages (**RQ3**)
- **RQ4**: general-purpose models w/o domain knowledge and fine-tuning are almost useless for the task, falling behind the simplistic tf-idf baseline. Our evaluation shows that clinical EL requires pre-training at least on the related biomedical corpora

Contacts

Anton Alekseev: anton.m.alexeyev@gmail.com
Elena Tutubalina: tutubalinaev@gmail.com
Artem Shelmanov: artemshelmanov@gmail.com

Code: