NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001313

AILABUD at the NTCIR-17 MedNLP-SC Task: Monolingual vs Multilingual Fine-tuning for ADE Classification

Beatrice Portelli
University of Udine
University of Naples Federico II
Italy
portelli.beatrice@spes.uniud.it

Alessandro Tremamunno
University of Udine
Italy
tremamunno.alessandro@spes.uniud.it

Simone Scaboro
University of Udine
Italy
scaboro.simone@spes.uniud.it

Emmanuele Chersoni
Hong Kong Polytechnic University
China
emmanuele.chersoni@polyu.edu.hk

Giuseppe Serra
University of Udine
Italy
giuseppe.serra@uniud.it

ABSTRACT

The AILAB team participated in the Social Media subtask of the NTCIR-17 MedNLP-SC Task. This paper reports our approach to solving the problem and discusses the official results. The presented model performs binary classification of the tweets and, given an UMLS term, determines whether it is present as an ADE in the tweet. Due to this design, it does not need an intermediate ADE extraction step, and it can be extended to new UMLS terms currently not present in the text. The base model used in the experiments is multilingual SapBERT, which was fine-tuned in a monolingual and multilingual setting. The best results were achieved by training the model on multilingual data.

KEYWORDS

Transformers, Multilingual Models, Monolingual Models, Medical Natural Language Processing, Named Entity Recognition, Adverse Drug Event

TEAM NAME

AILABUD

SUBTASKS

SM-ADE-JA
SM-ADE-EN
SM-ADE-DE
SM-ADE-FR

1 INTRODUCTION

The AILAB team participated in the Social Media subtask of the NTCIR-17 MedNLP-SC Task [10]. This paper reports our approach to solving the problem and discusses the official results.

Term normalization, a vital aspect of information processing systems, involves aligning various natural language expressions with specific entries in a dictionary or ontology. In the medical field, this task is crucial for connecting reported symptoms or adverse drug events with corresponding entries in medical ontologies like MedDRA [2] or UMLS [1]. However, this task is challenging due to the diverse range of natural language inputs, which can vary from casual social media content to formal medical/legal reports.

Although extensive research has been conducted on this problem for the English language in the past decade, there remains a lack of resources and methods for normalizing medical terms in other languages, especially in informal texts. To address this gap, efforts have been focused on developing multi-lingual zero-shot contrastive models. These models are pre-trained on large datasets of medical notes, research articles, and multi-lingual medical ontologies. While these embedding-based models showcase versatility, they still face challenges in effectively transferring their knowledge to informal language usage.

The objective of our team was to leverage multi-lingual zero-shot contrastive models, and fine-tune them in a mono- or multi-lingual setting to solve solve the ADE normalization challenge.

2 RELATED WORK

The study of medical term normalization in the past involved two main approaches: classification and ranking [12]. In the classification approach, neural models encode input terms to generate hidden representations and produce probability distributions over classes [6, 9]. However, this becomes challenging for extensive ontologies due to limited data. The ranking approach, on the other hand, focuses on prioritizing concepts based on their similarity to input terms [3, 7]. Systems are trained on binary classification tasks, where positive samples consist of term-concept pairs, while negative samples comprise term-concept pairs that do not match.

Advanced contrastive pre-training strategies have been applied in systems like BioSyn [8], CODER [12], SapBERT [4], and KRISSBERT [13], using UMLS to enhance medical embeddings in BERT-based models. These methods have significantly improved performance in various tasks, including zero-shot term normalization. While SapBERT employs self-alignment techniques, CODER maximizes similarities between positive term-term pairs and term-relation-term triples, achieving state-of-the-art results across various tasks, including zero-shot term normalization. Furthermore, KRISSBERT introduced an extensive pretraining procedure based on self-supervision, combining traditional masked language modeling with contrastive losses, which has proven highly effective for medical entity linking—a form of term normalization that utilizes the entire original context rather than just the extracted term.

However, most research has concentrated on English language resources, leading to a shortage of datasets and models for languages other than English, especially in handling informal texts, which present unique difficulties.

Noteworthy multi-lingual resources for social media data include those provided by the SMM4H workshop, [5, 11] which introduced small entity extraction and normalization datasets in Spanish, French, and Russian, although resources for non-English languages remain limited.

3 DATA

The task data consists of four datasets of tweets labeled for the presence of 22 ADEs. Each dataset contains the same tweets, but translated in different languages: Japanese, English, French, and German. Tweets can be matched to their translations using the sample ids provided in the datasets. The 22 ADEs are signified by their UMLS CUI (Concept Unique Identifier) and a string representing the CUI in the same language as the tweets (e.g., C0027497:nausea for the English dataset and C0027497:nausées for the French one). The training data contain 7964 samples.

For the following experiments we further split the training data into a training and a validation set with a 80:20 ratio. The split is performed on the list of sample ids, so for any given tweet all its translated versions belong either to the training or the validation set.

4 METHODS

This section describes our approach to solving the problem.

Given a text, we know that none, one or several ADEs might be contained within it.

We decided to develop two models: a binary model for ADE/noADE classification, to filter out samples that do not contain any ADEs, and a model for specific ADE classification.

4.1 ADE/noADE classification

To develop the model for binary ADE/noADE classification, we perform the following two steps, illustrated in Figure 1.

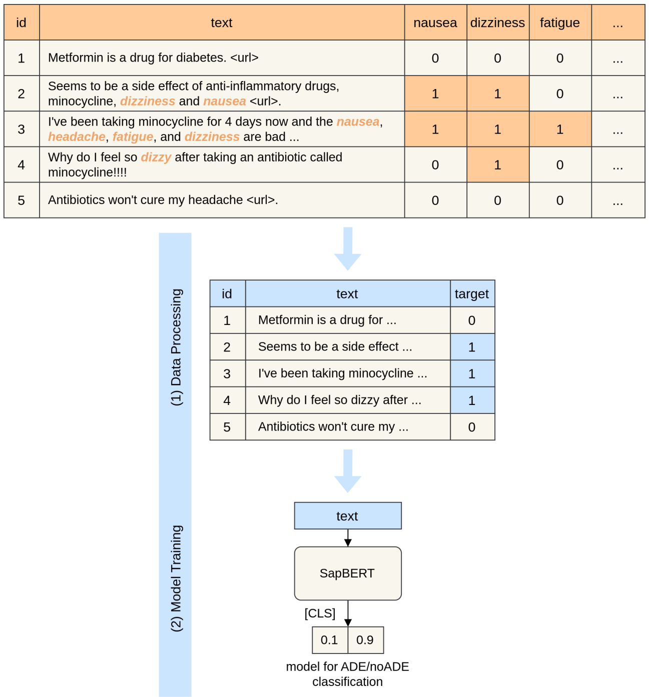
(1) Data Preprocessing. All train samples are converted to have binary labels. The label is 1 if at least one of the original 22 target labels is set to 1, and 0 otherwise.

(2) Model Training. A transformer-based model with a binary classification head is fine-tuned for classification on the training set.

The result is a trained model $\mathcal{M}_1$ to perform binary ADE/noADE classification.

4.2 Specific ADE classification

This is the most complicated part of the task, since a text might contain multiple ADEs and we are not aware of the text substring that represents each symptom (differently from what happens in other competitions such as SMM4H [11]). We decided to use the full text instead of extracting the possible ADE mentions from the tweet in order to avoid compounding the error of two models (extraction and normalization). Therefore, we device a flexible way to query the same text for the presence of multiple ADEs, framing it as a binary classification task.



id	text	nausea	dizziness	fatigue	...
1	Metformin is a drug for diabetes. <url>	0	0	0	...
2	Seems to be a side effect of anti-inflammatory drugs, minocycline, ***dizziness*** and ***nausea*** <url>.	1	1	0	...
3	I've been taking minocycline for 4 days now and the ***nausea***, ***headache***, ***fatigue***, and ***dizziness*** are bad ...	1	1	1	...
4	Why do I feel so ***dizzy*** after taking an antibiotic called minocycline!!!!	0	1	0	...
5	Antibiotics won't cure my headache <url>.	0	0	0	...

id	text	target
1	Metformin is a drug for ...	0
2	Seems to be a side effect ...	1
3	I've been taking minocycline ...	1
4	Why do I feel so dizzy after ...	1
5	Antibiotics won't cure my ...	0

Figure 1: Schema of the training procedure for the ADE/noADE classification model.

We perform the following two steps, illustrated in Figure 2.

(3) Data Preprocessing. All samples are converted to have binary labels. For each original text, we create 22 new samples, one for each ADE, where the sting of the ADE is attached in front of the text as follows: "<ADE> </s> <text>". The new samples are labeled as 1 if the specific ADE is contained in the text, and 0 otherwise. This system allows to query the text for the presence of multiple ADEs and is easily generalizable to new ADEs not seen during training.

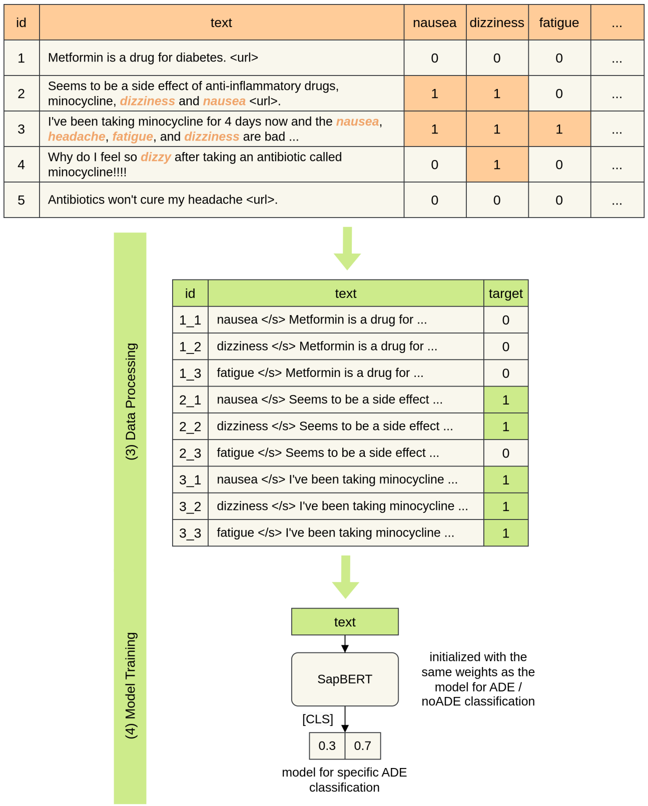
(4) Model Training. A transformer-based model with a binary classification head is fine-tuned for classification on the training set. The weights of the models are initialized with the ones of $\mathcal{M}_1$ (obtained at step (2)), so the model starts with a general understanding of the domain.

The result is a trained model $\mathcal{M}_2$ to perform specific ADE classification. At inference time, the predictions can be obtained by creating 22 samples for each new text, one for each ADE.

4.3 Monolingual and Multilingual Experiments

Since the same tweet is translated in multiple languages, we experiment with two settings:

- Monolingual experiment: we create separate models for each language ($\mathcal{M}_{2:\mathrm{ENG}}$, $\mathcal{M}_{2:\mathrm{FRE}}$, $\mathcal{M}_{2:\mathrm{GER}}$, $\mathcal{M}_{2:\mathrm{JAP}}$), where each one of them has only seen tweets in the respective language;



id	text	nausea	dizziness	fatigue	...
1	Metformin is a drug for diabetes. <url>	0	0	0	...
2	Seems to be a side effect of anti-inflammatory drugs, minocycline, *dizziness* and *nausea* <url>.	1	1	0	...
3	I've been taking minocycline for 4 days now and the *nausea*, *headache*, *fatigue*, and *dizziness* are bad ...	1	1	1	...
4	Why do I feel so *dizzy* after taking an antibiotic called minocycline!!!!	0	1	0	...
5	Antibiotics won't cure my headache <url>.	0	0	0	...

id	text	target
1_1	nausea </s> Metformin is a drug for ...	0
1_2	dizziness </s> Metformin is a drug for ...	0
1_3	fatigue </s> Metformin is a drug for ...	0
2_1	nausea </s> Seems to be a side effect ...	1
2_2	dizziness </s> Seems to be a side effect ...	1
2_3	fatigue </s> Seems to be a side effect ...	0
3_1	nausea </s> I've been taking minocycline ...	1
3_2	dizziness </s> I've been taking minocycline ...	1
3_3	fatigue </s> I've been taking minocycline ...	1

(3) Data Processing
(4) Model Training
text
SapBERT
initialized with the same weights as the model for ADE / noADE classification
[CLS]
0.3
0.7
model for specific ADE classification

Figure 2: Schema of the training procedure for the specific ADE classification model.

- Multilingual experiment: we create a single model that was trained on all texts ($\mathcal{M}_{2:\text{ALL}}$), regardless of the language, and use it to infer on all datasets.

4.4 Model Specifics and Hyperparameters

The base model we chose for the experiments is a multilingual SapBERT [4]. We selected this model because: it is pre-trained specifically on UMLS medical terms and it is pre-trained in a multilingual setting.

The following are the hyperparameters used for the models and the preprocessing steps. The hyperparameters remained the same between the first and second phases of the training, except for the number of epochs: in the first phase it was lowered to 3, because of an earlier loss convergence.

(2) Model Training.

- training epochs: 3
- learning rate: $2 * 10^{-5}$
- batch size = 32
- weight decay = 0.01

(3) Data Preprocessing. The number of potential training samples after preprocessing procedure increases from 6371 to 140162, the overwhelming majority of which belongs to class 0. This highly unbalanced datasets leads to several issues at training time. Therefore, we decided to limit the number of training samples in the following way: first, we remove 90% of the original samples that contain no ADEs, then we generate the 22 new samples for each tweet, and finally we remove 70% of the new samples that are labeled with 0. In this way we obtain a more balanced training set of around 18000 samples.

(4) Model Training.

- training epochs: 5
- learning rate: $2 * 10^{-5}$
- batch size = 32
- weight decay = 0.01

5 EXPERIMENTS

This section discusses the official results of the proposed models on the blind test set of the Task. In all the tasks, submission-1 was the monolingual model, while submission-2 was the multilingual model.

5.1 ADE/noADE classification

Table 1 reports the results for binary ADE/noADE classification. The multilingual $\mathcal{M}_{2:\text{ALL}}$ models always performs better than the monolingual models ($\mathcal{M}_{2:\text{ENG}}$, $\mathcal{M}_{2:\text{FRE}}$, $\mathcal{M}_{2:\text{GER}}$, and $\mathcal{M}_{2:\text{JAP}}$). The model $\mathcal{M}_{2:\text{ALL}}$ reaches its best performance on the English and Japanese datasets (0.83 and 0.82 F1 score). The monolingual models $\mathcal{M}_{2:\text{ENG}}$ and $\mathcal{M}_{2:\text{JAP}}$ are also the best-performing one among the monolingual ones.

Table 1: Results of the binary evaluation (ADE vs noADE) on the blind test set.

			ADE			noADE	
Dataset	Model	P	R	F1	P	R	F1
Japanese	$\mathcal{M}_{2:\text{JAP}}$	0.52	0.97	0.68	0.98	0.64	0.78
Japanese	$\mathcal{M}_{2:\text{ALL}}$	0.57	0.98	0.72	0.99	0.71	0.82
English	$\mathcal{M}_{2:\text{ENG}}$	0.54	0.95	0.69	0.97	0.67	0.80
English	$\mathcal{M}_{2:\text{ALL}}$	0.58	0.97	0.72	0.98	0.71	0.83
German	$\mathcal{M}_{2:\text{GER}}$	0.51	0.95	0.66	0.97	0.63	0.76
German	$\mathcal{M}_{2:\text{ALL}}$	0.56	0.96	0.70	0.98	0.69	0.81
French	$\mathcal{M}_{2:\text{FRE}}$	0.52	0.96	0.67	0.98	0.64	0.77
French	$\mathcal{M}_{2:\text{ALL}}$	0.55	0.97	0.70	0.98	0.69	0.81

5.2 Specific ADE classification

Table 2 reports the results per symptom class, calculating the performance for each symptom class and then averaging them. The multilingual model always performs better than the monolingual models on all metrics (macro precision, recall and F1-score). However, we can observe that all models have an extremely high recall (over 90%) and a low precision. We can therefore conclude that all models lack a precise understanding of which ADEs are present in the tweets and generate several false positives. A similar behavior can be observed also on the per-ADE metrics (Table 3), where the performance is calculated on the 0 and 1 labels regardless of the symptom class.

In both tables, the best performance is reached on the Japanese dataset, while the worst one is the performance on the French dataset. The same is true both for the monolingual and the multilingual models. Since the Japanese samples are the only ones which were manually annotated, the low performance on the French samples might be due to inconsistencies introduced during the Japanese-French translation.

Table 2: Results (macro average) per symptom class on the blind test set.

Dataset	Model	P	R	F1
Japanese	$\mathcal{M}_{2:\text{JAP}}$	0.48	0.93	0.61
Japanese	$\mathcal{M}_{2:\text{ALL}}$	0.53	0.96	0.67
English	$\mathcal{M}_{2:\text{ENG}}$	0.48	0.94	0.62
English	$\mathcal{M}_{2:\text{ALL}}$	0.52	0.95	0.65
German	$\mathcal{M}_{2:\text{GER}}$	0.44	0.92	0.58
German	$\mathcal{M}_{2:\text{ALL}}$	0.50	0.93	0.63
French	$\mathcal{M}_{2:\text{FRE}}$	0.45	0.94	0.59
French	$\mathcal{M}_{2:\text{ALL}}$	0.48	0.93	0.62

Table 3: Results per ADE label on the blind test set.

			0			1	
Dataset	Model	P	R	F1	P	R	F1
Japanese	$\mathcal{M}_{2:\text{JAP}}$	1.00	0.98	0.99	0.51	0.95	0.66
Japanese	$\mathcal{M}_{2:\text{ALL}}$	1.00	0.99	0.99	0.58	0.97	0.72
English	$\mathcal{M}_{2:\text{ENG}}$	1.00	0.98	0.99	0.47	0.94	0.63
English	$\mathcal{M}_{2:\text{ALL}}$	1.00	0.98	0.99	0.51	0.95	0.66
German	$\mathcal{M}_{2:\text{GER}}$	1.00	0.98	0.99	0.47	0.93	0.63
German	$\mathcal{M}_{2:\text{ALL}}$	1.00	0.98	0.99	0.54	0.94	0.69
French	$\mathcal{M}_{2:\text{FRE}}$	1.00	0.98	0.99	0.43	0.94	0.59
French	$\mathcal{M}_{2:\text{ALL}}$	1.00	0.98	0.99	0.48	0.94	0.64

5.3 Overall Performance

Overall, the best model considering exact match accuracy is the multilingual model $\mathcal{M}_{2:\text{ALL}}$ (see Table 4), which reaches 0.75 accuracy on Japanese, 0.71 on English and German, and 0.67 on French. As observed before, both monolingual and multilingual models reach their best performance on the Japanese dataset (originally annotated data) and their lowest performance on the French dataset (machine translation).

Table 4: Results of the exact match accuracy, where both binary classification and ADE classification must be correct to gain a point.

		Multilingual
Dataset	Monolingual	$\mathcal{M}_{2:\text{ALL}}$
Japanese	0.67	0.75
English	0.67	0.71
German	0.64	0.71
French	0.62	0.67

6 CONCLUSIONS

We developed a two-step method to solve the ADE classification task on social media data presented in the MedNLP-SC Social Media subtask. The presented model performs binary classification of the tweets and, given an UMLS term, determines whether it is present as an ADE in the tweet. Due to this design, it does not need an intermediate ADE extraction step, and it can be extended to new UMLS terms currently not present in the text. The base model used in the experiments is multilingual SapBERT, which was fine-tuned in a monolingual and multilingual setting. The best results were achieved by training the model on multilingual data.

The main limitation of the proposed approach is its high false positive rate (high recall, low precision). This could be caused by the lack of an ADE extraction step, which makes the tweets more difficult to process. Most of the samples contain more than one ADE, so this could create confusing signals and interactions during training. The base model (SapBERT) was also pre-trained with samples containing short UMLS terms only, not long complex sentences, so the domain shift might have been too severe. In the future, it would be interesting to add an ADE extraction module to measure its effect on the system, or experiments with other base models that were pre-trained on longer medical texts.

REFERENCES

[1] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research* 32 (02 2004), D267–70. https://doi.org/10.1093/nar/gkh061

[2] Elliot G Brown, Louise Wood, and Sue Wood. 1999. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety* 20, 2 (1999), 109–117.

[3] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. CNN-based Ranking for Biomedical Entity Normalization. *BMC Bioinformatics* 18, 11 (2017), 79–86.

[4] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proceedings of ACL-IJCNLP 2021*. 565–574.

[5] Arjun Magge, Ari Klein, Antonio Miranda-Escalada, Mohammed Ali Al-Garadi, Ilseyar Alimova, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima López, Ivan Flores, Karen O'Connor, Davy Weissenbacher, Elena Tutubalina, Abeed Sarker, Juan Banda, Martin Krallinger, and Graciela Gonzalez-Hernandez. 2021. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*. Association for Computational Linguistics, Mexico City, Mexico, 21–32. https://doi.org/10.18653/v1/2021.smm4h-1.4

[6] Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2019. Multi-task Character-level Attentional Networks for Medical Concept Normalization. *Neural Processing Letters* 49, 3 (2019), 1239–1256.

[7] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of ACL*.

[8] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3641–3650. https://doi.org/10.18653/v1/2020.acl-main.335

[9] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical Concept Normalization in Social Media Posts with Recurrent Neural Networks. *Journal of Biomedical Informatics* 84 (2018), 93–102.

[10] Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-17*. National Institute of Informatics (NII). https://doi.org/10.20736/0002001327

[11] Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena

Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the Seventh Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*. Association for Computational Linguistics, Gyeongju, Republic of Korea, 221–241. https://aclanthology.org/2022.smm4h-1.54

[12] Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. CODER: Knowledge-infused Cross-lingual Medical Term Embedding for Term Normalization. *Journal of Biomedical Informatics* 126 (2022), 103983. https://doi.org/10.1016/j.jbi.2021.103983

[13] Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Knowledge-Rich Self-Supervised Entity Linking. *CoRR* abs/2112.07887 (2021). arXiv:2112.07887 https://arxiv.org/abs/2112.07887