Improving Multi-lingual Medical Term Normalization to Address the Long-Tail Problem

Beatrice Portelli[1,2] portelli.beatrice@spes.uniud.it
Simone Scaboro[1] scaboro.simone@gmail.com
Giuseppe Serra[1] giuseppe.serra@uniud.it

[1]University of Udine, Udine, IT
[2]University of Naples Federico II, Naples, IT

Abstract

Medical term normalization involves the mapping of text to specific medical terms within a medical ontology. However, due to the vast number of possible medical terms and the scarcity of annotated datasets, this task becomes particularly challenging, especially for languages other than English, where the problem is further amplified. In this paper, we propose an approach to tackle this challenge by experimenting with the ontology pre-training (OP) method. We explore its potential for generalization across multiple languages. The core of this method lies in utilizing a large medical ontology, such as MedDRA, to generate synthetic samples in various languages to use during the model's pre-training. This augmentation technique aims to enhance the model's ability to generalize to classes that were not encountered during the fine-tuning process. To assess the effectiveness of our approach, we compare the performance of a robust zero-shot multilingual model with traditional fine-tuning, ontology pre-training, and their combined strategies. We experiment on three datasets for medical entity normalization belonging to different languages (English, French, and Russian), and analyze the effect of the presence/absence of the target language in the pre-training step. The results demonstrate the successful extension of ontology pre-training to multiple languages. We observe that multi-language pre-training significantly improves the baseline performance of models, enabling them to achieve strong performance without any loss when fine-tuned on new languages.

1 Introduction

Term normalization, a crucial task in information processing systems, involves the mapping of diverse natural language expressions to specific concepts within a dictionary or ontology. This process is particularly important in the medical domain, where it plays a significant role in associating reported symptoms or adverse events linked to drugs with corresponding entries in medical ontologies, such as MedDRA (Brown et al., 1999) or UMLS (Bodenreider, 2004). However, this task presents considerable challenges due to the wide variability of natural language inputs, ranging from informal social media and conversational transcripts to formal medical and legal reports. Additionally, the output concepts exhibit a high cardinality and follow a long-tail distribution, further complicating the normalization process.

While this problem has been extensively studied for the English language over the past decade, resources and methods for medical term normalization in other languages, especially on informal texts, remain scarce.

To address this issue, research efforts have been directed toward developing multi-lingual zero-shot contrastive models, pre-trained on large collections of medical notes, research articles, and multi-lingual medical ontologies. Although these embedding-based models demonstrate versatility, they lack the ability to effectively transfer their knowledge to more informal language usage.

A recently-introduced technique, called ontology pre-training (OP) (Portelli et al., 2022), has shown the potential to alleviate this challenge. OP utilizes a medical ontology to generate artificial samples, and has been shown to enhance the performance of term normalization models on English datasets with different writing styles. Our proposal is to extend this method to multiple languages and investigate the impact of including/excluding target languages during the pre-training phase on the downstream performance of the model.

In light of this objective, we aim to address the following research questions:

1. Is using OP in a different language more effective than using a zero-shot multilingual model?

2. What is the influence of incorporating multiple languages during OP on the overall effectiveness of the model?

3. Is using OP in a different language more effective than fine-tuning on the target language?

4. Can using OP in a different language serve as an effective starting point for further fine-tuning the model on the target language?

By investigating these research questions, we seek to contribute to the advancement of multilingual medical term normalization and facilitate more accurate and efficient medical information processing across languages.

2 Related Work

Medical term normalization has been widely explored as both a classification and ranking problem (Yuan et al., 2022). In the classification approach, neural architectures encode the input term into a hidden representation and output a distribution over classes (Tutubalina et al., 2018; Niu et al., 2019). However, scaling this approach to ontologies containing thousands of concepts becomes challenging due to the scarcity of comprehensive datasets.

On the other hand, the ranking approach aims to rank concepts based on their similarity to the input term (Li et al., 2017; Sung et al., 2020a). In this method, systems are trained on binary classification tasks, where positive samples consist of term-concept pairs, while negative samples consist of term-concept pairs that do not match. The raw output of the model is then used for ranking the concepts.

Recent advances have successfully combined classification and ranking approaches. For instance, Ziletti et al. (2022) proposed a system that integrates a BERT-based classifier (Devlin et al., 2019) with a zero/few-shot learning method to incorporate label semantics in input instances (Halder et al., 2020), leading to improved performance in single-model and ensemble settings.

Notably, novel contrastive pre-training strategies have been introduced in systems like BioSyn (Sung et al., 2020b), CODER (Yuan et al., 2022), SapBERT (Liu et al., 2021), and KRISSBERT (Zhang et al., 2021), which leverage UMLS to enhance medical embeddings in BERT-based models. While SapBERT employs self-alignment methods, CODER maximizes similarities between positive term-term pairs and term-relation-term triples, achieving state-of-the-art results on various tasks, including zero-shot term normalization. Moreover, KRISSBERT introduced an extensive pre-training procedure based on self-supervision and a combination of traditional masked language modeling with contrastive losses, which proved highly effective for medical entity linking, a type of term normalization that utilizes the full original context (instead of using only the extracted term).

While significant progress has been made, the focus has primarily been on English language resources, and there remains a considerable scarcity of datasets and models for languages other than English, especially regarding informal texts, which are the most challenging ones.

Some notable resources are the ones shared by SMM4H, (Weissenbacher et al., 2022; Magge et al., 2021) a yearly workshop that recently introduced small entity extraction and normalization datasets in Spanish, French, and Russian, but the resources for non-English languages are still limited. Another interesting addition is the MedNLP-SC social media task[1], which consists of adverse drug event detection and normalization from tweets in four languages: Japanese, English, French, and German. However, all the data are synthetic, as the Japanese tweets are generated using a language model and then automatically translated to the other languages. The Japanese samples are still annotated by humans, so the dataset might prove to be useful for future research.

In contrast, medical ontologies often have official translations available, offering an intriguing opportunity to explore the possibility of leveraging them to address similar tasks in non-English languages. Exploring multi-lingual normalization approaches could pave the way for more inclusive and effective medical information processing systems across languages.

3 Datasets

3.1 Ontology

We select MedDRA (Medical Dictionary for Regulatory Activities) (Brown et al., 1999) as a medical ontology, a subset of the UMLS (Bodenreider, 2004) ontology specialized in mapping diseases, symptoms, and medical procedures. In particular, we consider only the two lowest levels of the ontology: Preferred Terms (PT) and Lowest Level Terms (LLT). LLTs are the set of terms which are

[1] https://sociocom.naist.jp/mednlp-sc/

the closest to the ones used in everyday speaking (e.g., "muscle burning sensation"). PTs, on the other hand, are more formal terms, and are used to group several LLTs. For example, the LLTs "muscle burning sensation", "muscle ache", "tenderness muscle", and "localized muscle pain" are all grouped under the PT "myalgia".

The version of MedDRA we use contains 25,255 PTs and 51,109 LLTs. Each PT groups between 1 and 194 LLTs, with an average of 3 LLTs per PT.

We consider three languages for the ontology: English, French, and Russian. For each language we create a simplified tabular dataset that associates each LLT with its PT: `Eng`, `Fre`, and `Rus`.

3.2 Finetuning Datasets

We evaluate the models on three medical term normalization datasets that contain diseases, symptoms, or adverse events, all of which can be mapped to the MedDRA ontology. We selected datasets written in three different languages, namely English, French, and Russian, to better analyze the effects of the multilingual training.

CADEC (Karimi et al., 2015). Public dataset containing 1,250 English posts from the health forum "AskaPatient", containing user-reported adverse drug events mapped to a MedDRA PT/LLT. The language is informal.

Quaero (Névéol et al., 2014). Public dataset containing 2,498 French MEDLINE titles and 38 French EMEA documents, annotated for the presence of several categories of UMLS concepts. We consider only the DISORDER category.

RDRS (Sboev et al., 2022). Public dataset containing 3,821 Russian posts from the health forum "otzovik.com", containing user-generated medicine reviews, where medical entities are mapped to a MedDRA PT/LLT. The dataset also includes annotations for mentions of Medications. We consider the Disease and ADR (Adverse Drug Reaction) entities.

All datasets are preprocessed to contain only the medical entity mentions and their PT label. If the samples were labeled with an LLT in the original dataset, they are re-mapped to a PT using the MedDRA ontology to ensure an uniform output space containing only PT terms.

To test the generalization capabilities of the models, it is important to test them on different sets of unseen labels. For this reason, we create three random splits of train/test samples using a 60:40 proportion. Given a train and a test set, every test sample with label *pt* falls into one of the following categories:

- *seen*, if *pt* is present in the training set;
- *unseen*, if *pt* is *not* present in the training set.

The most important set of samples to measure the generalization capabilities of the models is *unseen*.

Name	**CADEC**	**Quaero**	**RDRS**
Language	English	French	Russian
Total samples	5,866	3,128	12,294
Train samples	3,535	1,876	7,376
Test samples	2,330	1,252	4,918
unseen %	4.76%	28.89%	4.37%
Unique PTs	443	1,154	929

Table 1: Statistics of the three datasets used, including the percentage of samples with *unseen* labels in the test set and the overall number of unique PTs.

Table 1 contains some statistics related to the datasets. We can see that Quaero is likely to be the most challenging dataset, as it contains the highest number of unique PTs (1,154) and the highest percentage of *unseen* samples in the test set (28.89%).

4 Models

For all the experiments, we used SapBERT multilingual (Liu et al., 2021), which is one of the best dataset-agnostic BERT-based models for medical term embedding. It was trained on the UMLS ontology (a super-set of MedDRA) using medical terms in over 20 languages. Tests on several multilingual term normalization datasets show that the models achieve promising results in all languages (Liu et al., 2021).

4.1 Baseline (Zero-shot)

As a baseline, we use SapBERT multilingual for zero-shot term normalization following its original paper.

First, we obtain a list of PT terms from the MedDRA ontology in one of the available languages (`Eng`, `Fre`, `Rus`). All the PTs are embedded using the model, generating a vocabulary of embedded terms. Given a sample s, the same embedding model is used to generate its embedding. The embedding of s is then compared with all the entries p in the pre-computed vocabulary. We select as

prediction the PT p that minimizes the cosine similarity with s.

$$CosineSim(s, p) = \frac{s \cdot p}{||s||\ ||p||}$$

Note that there is no need for the pre-computed vocabulary and the sample to belong to the same language.

4.2 Ontology Pre-training (OP)

Following Portelli et al. (2022), we use the simplified tabular ontology dataset (Section 3.1), which maps LLTs to PTs, to pre-train the model. A linear layer is added on top of the SapBERT model, with one output for each of the possible PT terms in the ontology. Give an ontology dataset (`Eng`, `Fre`, `Rus`), the model is trained to normalize each LLT to the correct PT. Differently from previous works, we also perform OP in multiple languages.

4.3 Fine-tuning (FT)

Fine-tuning is framed as a classification task. A linear layer is added on top of the SapBERT model, with one output for each of the possible PT terms in the ontology. Given a finetuning dataset (CADEC, Quaero, RDRS), the model is trained to output the correct PT for each input term.

Fine-tuning can be applied either directly to the baseline model (FT) or after the OP step (OP+FT).

4.4 Training and Evaluation Details

The hyperparameters used for training the models in the different configurations (FT, OP, and OP+FT) are reported in Table 2. All the experiments were carried out using an NVIDIA GeForce RTX 3090 GPU (24 GB).

The models are evaluated using accuracy, that is the percentage of samples labeled with the correct PT on the test set. The accuracy is computed on *all* samples, *seen* samples, and *unseen* samples. However, the accuracy on *unseen* samples is the most important metric to measure the generalization abilities of the models. Note that a sample is considered *unseen* when the model has not been *fine-tuned* on data containing its label. OP models have technically "seen" all possible output labels, however they have never encountered samples coming from the target (fine-tuning) dataset, which might have a different language and/or vocabulary to express the same concept, requiring a domain shift. All the reported metrics are an average over three runs.

Model		Num Epochs	Batch Size
FT	**CADEC**	10	16
	Quaero	10	4
	RDRS	10	16
OP	**1 language**	30	128
	2 languages	15	128
	3 languages	10	128
OP+FT	**CADEC**	+3	16
	Quaero	+3	4
	RDRS	+3	16

Table 2: Hyperparameters and training time for all the experiments.

	Test Dataset		
Vocabulary	**CADEC**	**Quaero**	**RDRS**
`Eng`	**45.72**	67.41	32.31
`Fre`	40.01	**69.60**	29.61
`Rus`	36.21	60.25	**36.20**

Table 3: Performance of the zero-shot models on *all* samples of the three datasets. Each model matches the samples in the dataset with the most similar term in the English, French or Russian MedDRA vocabulary. The best result for each dataset is bolded.

5 Results

5.1 Zero-shot

Table 3 reports the results for the zero-shot models on *all* samples of the datasets.

For each dataset, the best performance is achieved by the model which uses the pre-computed terms belonging to the same language as the dataset: `Eng` for CADEC, `Fre` for Quaero, and `Rus` for RDRS.

Another notable fact is that the datasets present different levels of difficulty: the best performance on Quaero is 69.60%, on CADEC it is 45.72%, and only 36.20% on RDRS. Since the samples in Quaero come form research articles and EMEA document, it contains more formal terms which are similar to the ones found in the medical ontology, and therefore easier to normalize. On the other hand, CADEC and RDRS contain informal terms written by internet users, and might contain typos or layman terms, making them extremely more challenging to normalize.

Lastly, we observe that even models using a different vocabulary language have a reasonable performance, confirming that SapBERT multilingual is a strong baseline model. For example, the best zero-shot model on CADEC is the En-

glish one, with 45.72% accuracy, but the French model reaches an accuracy of 40.01%, which is only 5 points apart. The performance of the English and French models are even closer on the French dataset Quaero, where the English model performs only 2.19 points worse than the French one.

5.2 Fine-tuning (FT)

Table 4 reports the performance of the fine-tuned models on the three datasets. As expected, fine-tuning leads to very high performances, almost doubling the accuracy of the models on CADEC and RDRS compared to the zero-shot models (e.g., from 36.20% to 76.08% on RDRS *all*). However, we can see that this is the effect of an extremely high performance on the *seen* partition only, while the performance on the *unseen* partition is 0 for all the models.

Interestingly, this leads to a degraded performance on the Quaero dataset: the performance of the zero-shot model on Quaero is 69.60%, while the FT model only reaches 57.67% on *all* samples. This is because Quaero has the highest percentage of *unseen* samples in the test set (as previously observed in Table 1).

Partition	CADEC	Quaero	RDRS
all	78.32	57.67	76.08
seen	82.23	81.10	79.55
unseen	00.00	00.00	00.00

Table 4: Performance of the fine-tuned models on the three datasets. The performance is reported for the samples with *seen* labels, *unseen* labels, and *all* samples.

5.3 Ontology Pre-training (OP)

Table 5 reports the performance of the OP models on the three datasets, considering *all* samples.

If we focus on the OP models that use only one language (first three rows), we observe that the accuracy on the test datasets is always higher than the one reached by the zero-shot models, even for the challenging Quaero dataset.

Using a model that is trained only in languages different from the one of the dataset (cells with white background) seems to lead to slightly worse performances compared with the zero-shot model. For example, `Eng` on Quaero and `Rus` on Quaero reach 64.59% and 49.52% respectively, both of which are lower than the 69.60% accuracy of the zero-shot model. The same is true for the RDRS dataset. CADEC seems to be an exception, as all the OP models perform better than the zero-shot one on this dataset.

If we combine different languages during OP, the results are mixed. Looking at the second set of rows in Table 5, we see that the OP model trained on `Eng + Fre` reaches the best performance on both CADEC and Quaero, surpassing the single-language models. This means that the English and French languages interact in a positive way, reinforcing the normalization capabilities of the model. On the other hand, the RDRS dataset does not benefit from the inclusion of other languages during OP. The French language seems to be the least-damaging one, as the models trained on `Fre + Rus` reach 60.80% on RDRS, compared with 61.04% using only `Rus`.

Using all three languages during OP, slightly lowers the performance on CADEC and Quaero (1-3 points), and lowers the performance on RDRS by 6 points.

	Test Dataset		
OP Model	**CADEC**	**Quaero**	**RDRS**
`Eng`	74.49	64.59	35.84
`Fre`	58.72	74.47	33.71
`Rus`	47.43	49.52	**61.04**
`Eng + Fre`	**75.52**	**76.89**	38.80
`Eng + Rus`	74.90	66.18	57.25
`Fre + Rus`	64.53	72.63	60.80
`Eng + Fre + Rus`	74.56	73.37	55.09
Zero-shot	45.72	69.60	36.20
FT	78.32	57.67	76.08

Table 5: Performance of the OP models on *all* samples of the three datasets. The best OP result for each dataset is bolded. A dark cell background means that the model was pre-trained on the language of the test dataset.

Table 6 reports the results of the same models on the *unseen* partition of the data.

All the OP models have higher accuracy than the FT models. Moreover, the accuracy of the models that have seen the target language during OP (darker cells) is close to the performance of the zero-shot models. For example, on the RDRS dataset, the models OP `Rus`, OP `Eng + Rus`, OP `Fre + Rus`, and OP `Eng + Fre + Rus`, all have an *unseen* accuracy within 6 points to the zero-shot model.

Considering the OP models which have never seen the target language (cells with white background), the ones which use multiple languages have a higher accuracy than the ones which use

	Test Dataset		
OP Model	CADEC	Quaero	RDRS
Eng	**57.14**	59.82	15.41
Fre	39.43	**69.59**	16.24
Rus	31.84	38.54	**33.64**
Eng + Fre	50.49	68.49	20.74
Eng + Rus	53.45	57.88	31.89
Fre + Rus	43.72	64.99	29.76
Eng + Fre + Rus	49.29	64.98	29.51
Zero-shot	45.72	69.60	36.20
FT	00.00	00.00	00.00

Table 6: Performance of the OP models on *unseen* samples of the three datasets. The best OP result for each dataset is bolded. A dark cell background means that the model was pre-trained on the language of the test dataset.

a single language. For example, on the RDRS dataset, OP Eng + Fre has an accuracy of 20.74%, higher than both OP Eng (15.41%) and OP Fre (16.24%).

Figure 1 (Appendix A) summarizes all the results, allowing us to visually compare the performance of the OP and FT models on the *seen*, *unseen*, and *all* data partitions. The first model on the x-axis is always the FT model, while the horizontal line represents the performance of the zero-shot baseline model.

5.4 Ontology Pre-training + Fine-tuning (OP+FT)

Finally, we analyze the OP+FT models. Tables 7 and 8 report the metrics of the OP+FT models on the *all* and *unseen* data partitions respectively.

First, we focus on the ones that have seen the target language during OP (cells with darker background). The performance of the OP+FT models on *all* and *unseen* samples is higher than FT models. The highest performance is always achieved by the models which have only seen one language during OP. However, the gap between single-language and multi-language models is reduced compared to 5. Indeed, the model that uses all three languages during OP performs only 1 point lower than the single-language models on *all* samples (e.g., 78.66% vs 79.28% on RDRS). The performance on the *unseen* samples is lower than the one reached by the OP models (Table 6) and the zero-shot models, because the fine-tuning has caused the models to partially forget information about the *unseen* classes.

If we focus on the models that have never seen the target language during OP (cells with white background), we can see that the performance on *all* samples is always higher than the one reached by the FT model. This means that any OP model can be used as an effective starting point to fine-tune a dataset-specific model, even if the language of the new dataset is different from the one it has seen during pre-training. As regards the performance on the *unseen* samples, it is lower than the one achieved by a zero-shot model. However, it seems that models pre-trained on multiple languages can help reach a higher accuracy on *unseen* samples, compared to single-language models, as seen for the OP models.

Figure 2 (Appendix A) summarizes all the results, allowing us to visually compare the performance of the OP+FT and FT models on the *seen*, *unseen*, and *all* data partitions.

	Test Dataset		
OP Model	CADEC	Quaero	RDRS
Eng	**85.01**	77.64	77.96
Fre	83.13	**80.83**	77.55
Rus	82.57	71.33	**79.28**
Eng + Fre	84.60	80.67	77.93
Eng + Rus	84.44	75.91	78.91
Fre + Rus	82.84	78.30	78.64
Eng + Fre + Rus	84.31	79.10	78.66
Zero-shot	45.72	69.60	36.20
FT	78.32	57.67	76.08

Table 7: Performance of the OP+FT models on *all* samples of the three datasets. The best OP+FT result for each dataset is bolded. A dark cell background means that the model was pre-trained on the language of the test dataset.

	Test Dataset		
OP Model	CADEC	Quaero	RDRS
Eng	**35.75**	40.46	05.09
Fre	19.15	**50.13**	04.13
Rus	16.20	22.12	**12.91**
Eng + Fre	31.76	49.94	07.45
Eng + Rus	30.32	35.20	11.52
Fre + Rus	20.72	43.13	10.65
Eng + Fre + Rus	28.80	43.77	12.06
Zero-shot	45.72	69.60	36.20
FT	78.32	57.67	76.08

Table 8: Performance of the OP+FT models on *unseen* samples of the three datasets. The best OP+FT result for each dataset is bolded. A dark cell background means that the model was pre-trained on the language of the test dataset.

6 Conclusions

We dealt with the task of medical entity normalization and the problem of generalizing over long-tail distributions. We tested the effectiveness of the ontology pre-training (OP) method in a multilingual setting, combining OP in multiple languages with FT.

We answered our initial research questions with the following take-home messages:

1. **Is using OP in a different language more effective than using a zero-shot multilingual model?**
   As seen in Section 5.3, generally speaking, using an OP model trained on a single language (different from the target language) will lead to lower performance than using a strong zero-shot multilingual model. The more the OP language is different, the more performance will be lost.

2. **What is the influence of incorporating multiple languages during OP on the overall effectiveness of the model?**
   As seen in Section 5.3, in general, using multiple languages during OP will lead to performances higher than zero-shot models. However, some languages interact better than others (e.g. English and French boost each other's performance better than English and Russian). This is probably due to the fact that some languages are more similar than others. Increasing the number and diversity of languages during OP could lead to better general performances.

3. **Is using OP in a different language more effective than fine-tuning on the target language?**
   As seen in Section 5.3, using OP in a different language always leads to higher performance on the *unseen* partition, compared with FT models. Using OP with multiple languages (all different from the target language) further improves the performance on *unseen* samples, without much degradation for the *seen* samples.

4. **Can using OP in a different language serve as an effective starting point for further fine-tuning the model on the target language?**
   As seen in Section 5.4, yes, a model pre-trained with OP in one or more languages different from the target one can be further fine-tuned on the target language. The performance on both *seen* and *unseen* samples is higher than a model which only uses FT, and close to the one of a model which has seen the target language during OP. The slight loss in performance on *unseen* samples (even when using the same language) is a well-known "forgetting" phenomenon caused by performing consecutive fine-tuning procedures on the same model. It would be interesting to explore methods to alleviate the forgetting phase (e.g., using continual learning or other strategies to prevent catastrophic forgetting (McInerney et al., 2021)).

In general, we conclude that the models produced from the OP method benefit from the use of multilingual training data. The models created by a multilingual OP training have a good performance across different languages and can be used as a starting point to fine-tune models in different languages which they were not trained on.

Limitations

The experiments were conducted using a single pre-trained model (SabBERT). Adding a variety of pre-existing pre-trained models would further support the results found in the study.

The experiments were only performed using three languages, English, French, and Russian, two of which have higher affinity between each other. It would be interesting to expand the analysis to other different languages, such as Japanese and German, for which the medical ontology is available. The major setback in this direction is the scarcity of annotated data for the medical entity normalization task in languages other than English, especially if we consider non-formal texts (e.g., user feedback and reviews), which are the most challenging and interesting ones.

References

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32:D267–70.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-aware Representation of Sentences for Generic Text Classification. In *Proceedings of COLING*.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chenchen Wang. 2015. Cadec: A Corpus of Adverse Drug Event Annotations. *Journal of Biomedical Informatics*, 55:73–81.

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):79–86.

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*, pages 565–574.

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*, pages 21–32, Mexico City, Mexico. Association for Computational Linguistics.

Denis Jered McInerney, Chris (Luyang) Kong, Kristjan Arumae, Byron Wallace, and Parminder Bhatia. 2021. Kronecker factorization for preventing catastrophic forgetting in large-scale medical entity linking. In *NeurIPS 2021 Workshop on Machine Learning in Public Health*.

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):1239–1256.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.

Beatrice Portelli, Simone Scaboro, Enrico Santus, Hooman Sedghamiz, Emmanuele Chersoni, and Giuseppe Serra. 2022. Generalizing over long tail concepts for medical term normalization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8580–8591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Sboev, Sanna Sboeva, Ivan Moloshnikov, Artem Gryaznov, Roman Rybka, Alexander Naumov, Anton Selivanov, Gleb Rylkov, and Vyacheslav Ilyin. 2022. Analysis of the full-size russian corpus of internet drug reviews with complex ner labeling using deep learning neural networks and language models. *Applied Sciences*, 12(1).

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

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020b. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

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*, 84:93–102.

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*, pages 221–241, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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*, 126:103983.

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.

Angelo Ziletti, Alan Akbik, Christoph Berns, Thomas Herold, Marion Legler, and Martina Viell. 2022. Medical coding with biomedical transformer ensembles and zero/few-shot learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 176–187, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

A Extended Results

Figure 1 summarizes all the results for the OP models, allowing us to visually compare the performance of the OP and FT models on the *seen*, *unseen*, and *all* data partitions. The first model on the x-axis is always the FT model, while the horizontal line represents the performance of the zero-shot baseline model.

Figure 2 reports the same results for all the OP+FT models.

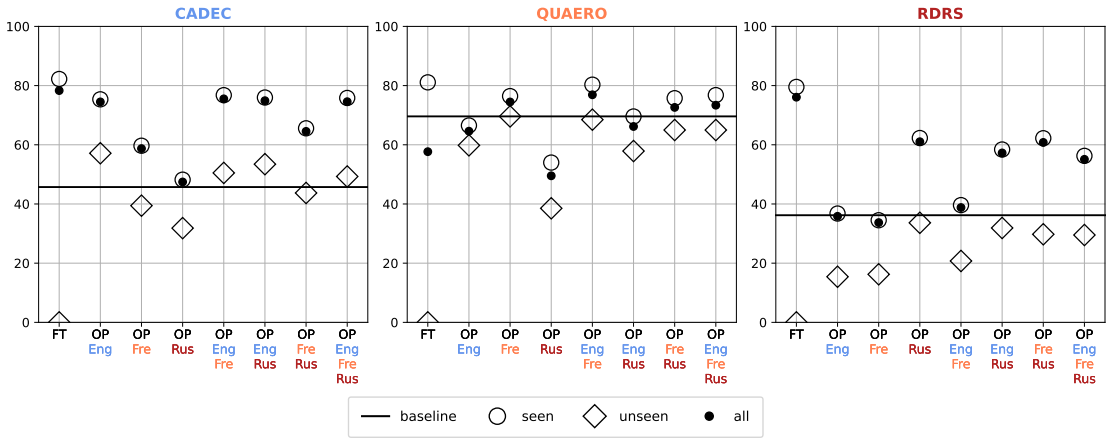

Figure 1: Performance of the FT and OP models on the three datasets, for each data partition (*seen*, *unseen* and *all*). The baseline is the zero-shot model in the target language.

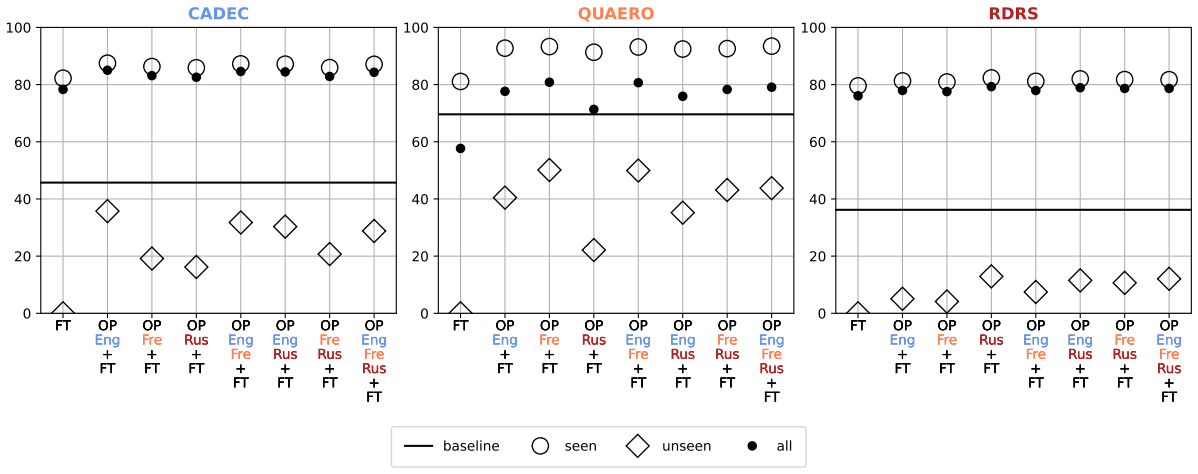

Figure 2: Performance of the FT and OP+FT models on the three datasets, for each data partition (*seen*, *unseen* and *all*). The baseline is the zero-shot model in the target language.