Give it a shot: Few-shot learning to normalize ADR mentions in Social Media posts

Manolis Manousogiannis
myTomorrows
Delft University of Technology
m.manousogiannis@mytomorrows.com

Sepideh Mesbah
Delft University of Technology
s.mesbah@tudelft.nl

Selene Baez Santamaria
myTomorrows
s.baez@mytomorrows.com

Alessandro Bozzon
Delft University of Technology
a.bozzon@tudelft.nl

Robert-Jan Sips
myTomorrows
r.sips@mytomorrows.com

Abstract

This paper describes the system that team MYTOMORROWS-TU DELFT developed for the 2019 Social Media Mining for Health Applications (SMM4H) Shared Task 3, for the end-to-end normalization of ADR tweet mentions to their corresponding MEDDRA codes. For the first two steps, we reuse a state-of-the-art approach, focusing our contribution on the final entity-linking step. For that we propose a simple Few-Shot learning approach, based on pre-trained word embeddings and data from the UMLS, combined with the provided training data. Our system (relaxed F1: 0.337-0.345) outperforms the average (relaxed F1 0.2972) of the participants in this task, demonstrating the potential feasibility of few-shot learning in the context of medical text normalization.

1 Introduction

Team MYTOMORROWS-TU DELFT participated in subtask 3 of the 2019 Social Media Mining for Health Applications (SMM4H) (Davy Weissenbacher, 2019) workshop, which is an end-to-end task. The goal is, given a tweet, to 1) automatically classify tweets containing an adverse drug reaction mention; 2) extract the exact ADR mention; 3) normalize the extracted ADR to its corresponding Medical Dictionary for Regulatory Activities (MEDDRA) code. The task is evaluated based on strict and relaxed F-score, precision and recall.

From an NLP perspective, this task poses a significant challenge as there is a large gap between the informal language used in social media and the formal medical language. Moreover, there is an absence of large annotated datasets, and datasets which are available often suffer from class imbalance. Illustrating this, Figure 1 provides an overview of the number of samples per class in the SMM4H task 3 dataset.

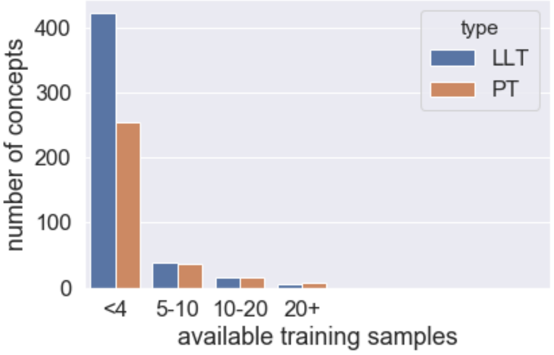

Figure 1: Available training samples per the medical concept present in the training data

Our end-to-end system consists of existing state-of-the-art for the first two steps. We focus our efforts on the third -normalization- step, which we formulate as a Few-Shot Learning problem (FSL), following the definition by Wang and Yao (Wang and Yao, 2019). In the following sections, we describe (1) the datasets that we worked on, (2) our approach in more detail and finally (3) our results and conclusions.

2 Data

2.1 Datasets

With the three subtasks, three manually annotated datasets were provided. All datasets contain tweets containing an ADR (positive) and without an ADR (negative). A brief overview of these datasets is provided in Table 1, but for more context we refer to (Davy Weissenbacher, 2019).

2.2 Preprocessing

The provided dataset for subtask 3 consists of ADR mentions, annotated with their corresponding MEDDRA code. In the hierarchy[1] of MEDDRA,

[1] https://www.meddra.org/how-to-use/basics/hierarchy

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 114–116
Florence, Italy, August 2, 2019.

Task	Training data	
	#Positives	*#Negatives*
1	2374	23298
2	1212	1155
3	1212	1155

Table 1: Statistics of the training data used for task 1, 2 and 3

one Preferred Term (PT) is linked to one or more Lower Level Terms (LLTs) which are more specific descriptions of the related concept.

The provided dataset contains a mix of PTs and LLTs, mapping the 1212 ADR mentions to more than 500 different codes. Observing that the evaluation of the workshop task is performed on PT level, we map all annotations to the corresponding PT, as a preprocessing step. After this preprocessing step, the 1212 training mentions are mapped to 319 MEDDRA codes. Figure 1 provides an overview of the class distribution before and after preprocessing.

2.3 Prior Knowledge

In the training set for subtask 3, 149 out of the 319 MEDDRA codes that are present in the dataset (46.7%) have just one available training sample, while 254 (79.6%) have less than five training samples. To deal with the scarcity of samples, we create a prior knowledge dataset considering the 319 MEDDRA PTs in the training data. This dataset consists of the preferred names provided by the MEDDRA vocabulary and their corresponding preferred names in the Consumer Health Vocabulary (CHV), as mapped by the UMLS. The resulting dataset cointains 1,854 preferred names for the 319 MEDDRA codes.

3 Method

Our contributions focus on the normalization step, linking ADRs to their corresponding MEDDRA code. However, to be able to perform an end-to-end evaluation, we use existing state-of-the art techniques for subtask 1 (Sarker and Gonzalez, 2015) and 2 (Cocos et al., 2017), which we train on the workshop datasets [2].

The state-of-the-art approach for medical concept normalization in user-generated text is deep-neural networks (Limsopatham and Collier, 2016) which outperform traditional methods, when sufficient training data are available.

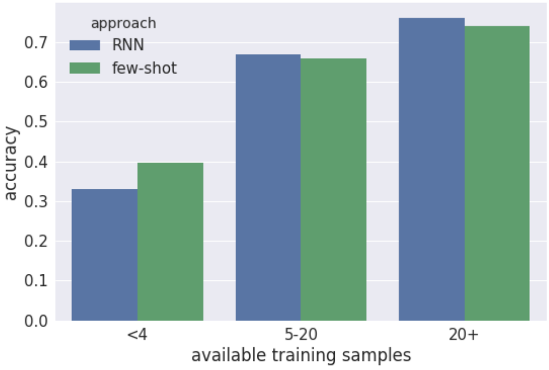

Figure 2: Accuracy per number of training samples.

We trained both the CNN and RNN described by (Limsopatham and Collier, 2016) on the dataset for task 3, finding that the RNN has the best performance. On closer observation (and not surprisingly), we found that the accuracy of the RNN drops when fewer samples are available in the training data, as depicted in figure 2.

To deal with this drop in performance, we propose an embedding-based classifier that compares the ADR extracted mention to its 1-Nearest Neighbour on a vector space containing a) representations of the ADR mentions in the training data and b) representations of the prior knowledge dataset. Our intuition is that the embedding-based binary classifier would perform better on classes with a low number of samples, whereas an RNN would perform well on classes with higher sample numbers.

To create our embedding-based classifier we employ the pretrained Google News Word2Vec model (Mikolov et al., 2013). Using this model, we create vector representations for the ADR mentions in our training data[3]. Similarly we create vector representations for the mentions gathered in our prior knowledge dataset. At test time, we employ the same Word2Vec model to create a vector representation of the unseen ADR mention. Using a 1-Nearest Neighbour (with cosine similarity as distance metric), we then select the corresponding MEDDRA concept. Figure 2 shows that this model indeed seems less sensitive to low sample numbers.

[2] For task 1, we trained using the suggested settings, assigning 3:1 class weight favouring the ADR class. For task 2, we trained using the pre-trained-fixed setting.

[3] for mentions of more than one token we added the vectors

Technique	**Relaxed**			**Strict**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
RNN	*0.318*	*0.337*	*0.327*	*0.232*	*0.246*	*0.239*
FSL	**0.336**	**0.355**	**0.345**	**0.237**	**0.252**	**0.244**
RNN+FSL (1)	*0.328*	*0.347*	*0.337*	*0.23*	*0.244*	*0.237*
RNN+FSL (2)	*0.331*	*0.35*	*0.34*	*0.235*	*0.249*	*0.242*
Task 3 AVG	*0.29*	*0.311*	*0.297*	*0.205*	*0.224*	*0.211*

Table 2: Relaxed and strict Precision/Recall/F-score for RNN, FSL, RNN+FSL (1) and (2) and the average score of all the participated team in task 3 (Task 3 AVG)

For our experiments, we use 4 systems: (1) RNN: the RNN proposed by (Limsopatham and Collier, 2016), trained on the both prior knowledge and the training set (which provides the best performance), (2) FSL: our 1-NN based on a combination of prior knowledge and the training set, (3) RNN+FSL (1): an ensemble of the RNN trained on only the training set and the FSL based on training + prior knowledge, and (4) RNN+FSL (2): an ensemble of the RNN trained on the training set and prior knowledge and the FSL based on training + prior knowledge. For our ensembles, we trust the model with the highest confidence (we used the cosine similarity for the 1-NN model to represent confidence) in case of disagreement.

4 Results

Our results are summarized in Table 2. Despite the fact that the RNN+FSL performed better in our development set, it did not generalize in the test data. On the test and evaluation data, FSL outperformed all the other techniques and achieved a 0.345 relaxed F-score and a 0.244 strict F-score which are above the average performance achieved in this task by all participants (i.e. Task 3 AVG).

5 Conclusions

In this paper, we describe our approach in subtask 3 of the SMM4H shared task for normalization of Adverse drug reaction mentions in Twitter posts. Our few-shot learning approach performs above the average in this task and hence we believe it to be a promising approach in cases where the amount of training data is limited.

As future work, we will focus on the discrimination between the ADRs that belong to one of the 'commonly seen cases' (classes with sufficient training data) from the 'rare cases' (classes with insufficient training data). This will allow us to efficiently combine a deep neural network with a few-shot learning approach into a more robust system that successfully links ADR tweet mentions into its MEDDRA codes.

References

Anne Cocos, Alexander G Fiks, and Aaron J Masino. 2017. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821.

Arjun Magge Ashlynn Daughton Karen O'Connor Michael Paul Graciela Gonzalez-Hernandez. Davy Weissenbacher, Abeed Sarker. 2019. Overview of the fourth social media mining for health (smm4h) shared task at acl 2019. in proceedings of the 2019 acl workshop smm4h: The 4th social media mining for health applications workshop shared task.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1014–1023.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Yaqing Wang and Quanming Yao. 2019. Few-shot learning: A survey. *arXiv preprint arXiv:1904.05046*.