Biomedical Concept Normalization

Xuan Wang (xwang174@illinois.edu)
Fei Ling (fling2@illinois.edu)

Introduction

Concept normalization is also referred to as entity linking in general domain. Biomedical concept normalization is a follow-up task of biomedical named entity recognition (BioNER). After extracting and typing entity mentions from the text with BioNER tools, concept normalization tools will link the entity mentions to a concept identifier in a controlled vocabulary. For example, in Figure 1, the blue entity mentions (e.g., gentamicin sulfate) are the chemical entities and the yellow entity mentions (e.g., renal failure) are the disease entities recognized by a BioNER tool. The controlled vocabulary used here is Medical Subject Headings (MeSH). The goal of biomedical concept normalization is to link the entity mention "gentamicin sulfate" to the MeSH identifier "D005839: Gentamicins" on the concept ontology. Biomedical concept normalization is important since it can group synonym entity mentions together and produce a more structured output for downstream applications, such as relation extraction and knowledge base completion.

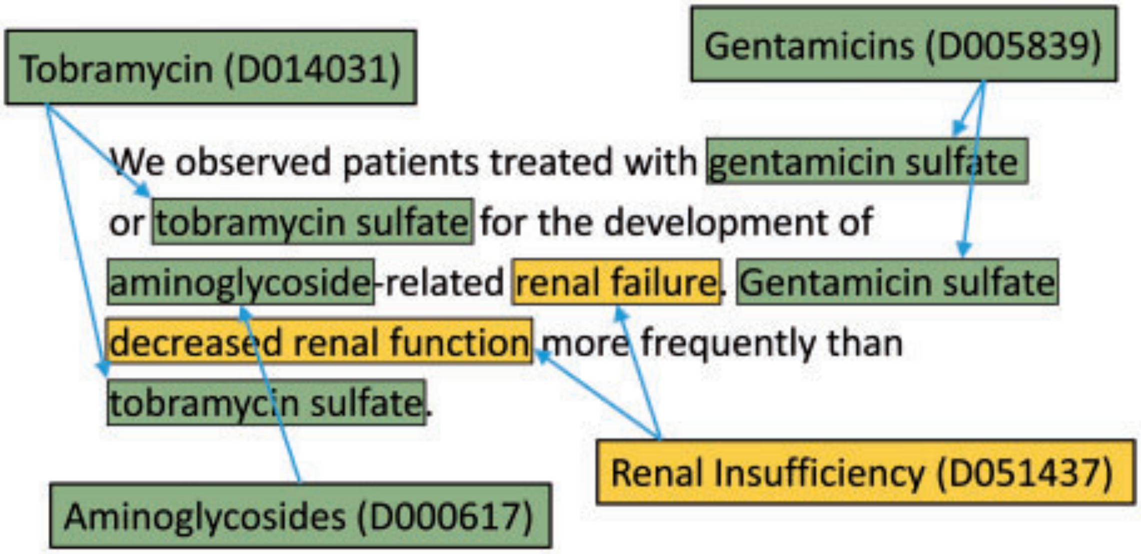

Tobramycin (D014031)
Gentamicins (D005839)
We observed patients treated with gentamicin sulfate
or tobramycin sulfate for the development of
aminoglycoside-related renal failure. Gentamicin sulfate
decreased renal function more frequently than
tobramycin sulfate.
Renal Insufficiency (D051437)
Aminoglycosides (D000617)

FIGURE 1. AN EXAMPLE OF BIOMEDICAL CONCEPT NORMALIZATION ON MESH CONCEPT ONTOLOGY.

Related work

Previous work often deal with biomedical concept normalization as a separate task of biomedical named entity recognition. For example, tmChem [1] is a tool for chemical entity recognition and normalization. It first performs entity recognition with an ensemble model consisting of two conditional random field (CRF) models. Then the output entity mentions are used for a lexicon look-up that links those entity mentions to concept identifiers in MeSH and ChEBI. The other example is DNorm [2], which is a tool for disease entity recognition and normalization. It also first performs entity recognition with a CRF model based on rich features. They use a pairwise learning to rank model to link the recognized entity mentions to the NCBI disease concept ontology. The above two models are the baseline models for chemical and disease entity recognition and normalization. The major problem with these models is that they treat entity recognition and concept normalization as two separate tasks, where the errors produced by the NER model will be propagated down to the next step of concept normalization. A better model would be to build an end-to-end system that jointly model the entity recognition and concept normalization tasks.

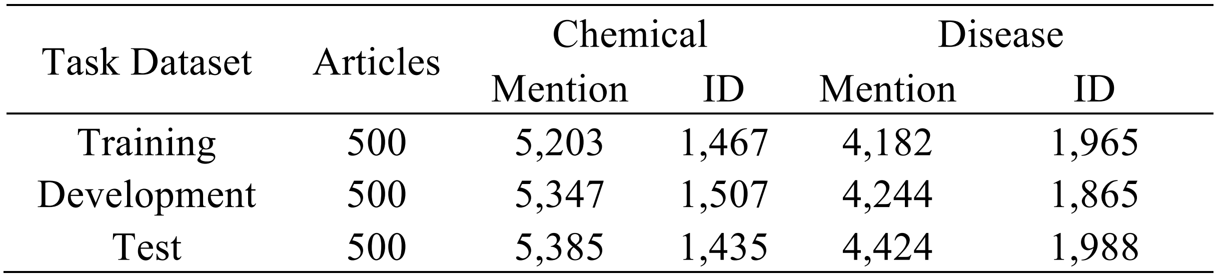
A recent work, TaggerOne [3], deals with the above problem by jointly model entity recognition and concept normalization with a semi-Markov model. It achieves better performance and is currently the state-of-the-art model for biomedical concept normalization. However, the features used in TaggerOne are simple linear transformations of the TF-IDF vectors of the entity mentions and concepts on the ontology. An improvement could be to use the feature representations learned from a neural network model with non-linear transformations to see if the performance could be further improved.

Method

Our end goal is to build a joint neural network model for biomedical named entity recognition and concept normalization. In this study, we re-implemented two baseline methods: dictionary look-up and pairwise learning to rank, to compare with previous studies on our new dataset. We focus on testing the performance of the concept normalization methods, so we use the gold standard entity mentions provided in the dataset.

Dataset

The dataset we used is BioCreative V CDR, which is a corpus for chemical and disease concept normalization. The detailed corpus statistics is shown in Table 1. There are 5000 PubMed articles used in each of the training, development and test set, respectively. This corpus includes two entity types: chemical and disease. There are around 5000 chemical entity mentions and 4000 disease entity mentions in each of the training, development and test set, respectively. The concept ontology we used is MeSH 2018.



Task Dataset	Articles	Chemical		Disease	
		Mention	ID	Mention	ID
Training	500	5,203	1,467	4,182	1,965
Development	500	5,347	1,507	4,244	1,865
Test	500	5,385	1,435	4,424	1,988

TABLE 1. OVERALL STATISTICS OF THE BIOCREATIVE V CDR CORPUS.

Dictionary look-up

The first method we tried is a simple dictionary look-up. For each concept identifier on MeSH, we use a list of all the synonyms under that identifier as its ontology concepts. We first normalize all the entity mentions and the concepts in the ontology to lowercase. Then we perform a stringent string matching to link the entity mentions in text to the MeSH ontology. This method is easy to implement and works quite robust and fast. We also implemented an API for users to have a try on this method.

Pairwise learning to rank

The concept normalization is not a standard classification problem. For example, there are around 270,000 concept identifiers as class labels in MeSH, and it is impossible for the human labeled training corpus to cover all the possible class labels. A better way is to model concept normalization as a pairwise ranking problem. The true label should always be ranked higher than the other labels in the ontology. During prediction, the label ranked highest will be the predicted class label (concept identifier) for the input entity mention.

We use a margin ranking perceptron model as our pairwise learning to rank model. The input feature vectors are the TF-IDF vectors of the entity mentions, *m*, and the ontology concepts *n*. The true label for an entity mention *m* is n^+, and all the labels in the ontology except n^+ are n^-. We train a weight matrix *W* so that $m^TWn^+ > m^TWn^-$.

The margin ranking loss function is as follows:

$$W = \underset{W}{\operatorname{argmin}} \sum_{m} \sum_{n^+} \sum_{n^-} max\big(0, 1 - m^T W n^+ + m^T W n^-\big)$$

We use the stochastic gradient descent (SGD) algorithm for optimization. We adopted the same parameter setting for the learning rate in DNorm. Since the MeSH ontology contains around 270,000 concept identifiers, the feature space is too large and it is hard to effectively train the weight matrix. We perform a feature dimension reduction with truncated-SVD to reduce the feature dimension to 100. The large number of concept identifiers on MeSH also lead to too many negative examples during training. We perform a negative sampling to sample 1000, 100 or 25 negative samples as n^- for each iteration. Due to the time limit, we currently only finished the training with 25 negative samples.

Results

The results of the micro-average precision, recall and f1 scores are shown in Table 2. The dictionary look-up method achieves a precision, recall and f1 score of 0.39, which indicates that there are still a lot of new entity mentions that have not been included in the MeSH ontology.

Our current implementation of the pairwise learning to rank model didn't achieve an expected performance. The precision, recall and f1 score are all close to zero. One major reason could be that, due to the time limit, we currently only finished training the weight matrix with 25 negative samples. However, in MeSH, there are around 270,000 to be ranked and should be used as negative examples. Our current parameters may not receive enough training.

	Precision	Recall	F1
Dictionary look-up	0.39	0.39	0.39
Pairwise learning to rank	-	-	-

TABLE 2. MICRO-AVERAGE SCORES OF THE TWO METHODS ON THE CDR CORPUS.

Discussion

We need to further improve the pairwise learning to rank model training. Currently, we are using all the 270,000 concept identifiers in the MeSH ontology as candidate labels. This label space may be too big to get a good ranking model. One improvement is to reduce the label space by only selecting those labels that are most related to our dataset, e.g., the chemical and disease labels. The other improvement is to increase the number of negative samples used during the training. We have also tried negative sampling with 1000 or 100 samples, but the training is not finished yet due to the time limit.

The next step is to build a joint neural network model for biomedical named entity recognition and concept normalization. We plan to use word embeddings as input into a neural network model (e.g., Bi-LSTM) to learn a better set of entity representations. These representations can be further feed into a CFR layer for named entity recognition and also our margin ranking perceptron model for concept normalization.

References

1. Leaman, Robert, Chih-Hsuan Wei, and Zhiyong Lu. "tmChem: a high performance approach for chemical named entity recognition and normalization." Journal of cheminformatics 7, no. 1 (2015): S3.
2. Leaman, Robert, Rezarta Islamaj Doğan, and Zhiyong Lu. "DNorm: disease name normalization with pairwise learning to rank." Bioinformatics 29, no. 22 (2013): 2909-2917.
3. Leaman, Robert, and Zhiyong Lu. "TaggerOne: joint named entity recognition and normalization with semi-Markov Models." Bioinformatics 32, no. 18 (2016): 2839-2846.