Biomedical Entity Normalization based on Pre-trained Model with Enhanced Information

Lu Fang, Yiling Cao, and Zhongguang Zheng

Fujitsu R&D Center Co., Ltd. Beijing 100022, China
{fanglu,caoyiling,zhengzhg}@fujitsu.com

Abstract. Biomedical entity normalization, which links entity mentions in biomedical texts to their corresponding standard concepts in a knowledge base(KB) or an ontology is an important task in biomedical text mining. A prevalent solution is to generate the most similar concepts, and then rank those concepts with semantic models. Herein, to improve the performance of candidate concepts ranking for entity normalization, we rank the candidates by fine-tuning the domain-specific pre-trained BioBERT model and enhancing the representation information with entity mentions and candidates. We have achieved significant improvement over the state-of-the-art method on the Bacteria Biotope data of BioNLP-OST19[1] .

Keywords: Biomedical Entity Normalization · KB · BioBERT

1 Introduction

Mapping entity mentions in texts to a certain standard knowledge base (KB) or an ontology is a fundamental task, which can link the unstructured text to a structured dataset. Ambiguity and variation are the main challenges of this task. Unlike in the general domain, variation is much more common than ambiguity in the biomedical domain. Therefore, a variety of methods have been proposed to deal with this challenge, including rule based methods [1], machine learning and deep learning based methods [2, 3].

Recently, pre-trained models have been applied to many NLP tasks in the biomedical domain such as named entity recognition and relation classification tasks, resulting in significant improvements. BioBERT [4], which is based on the BERT model and pre-trained on large-scale biomedical articles, is applied to many biomedical NLP tasks to improve state-of-the-art performance. But few researchers have used the models for entity normalization. In this paper, we propose a biomedical entity normalization approach by fine-tuning the BioBERT model. We enhance the representation information by using the embedding of the special first token as well as the embeddings of the entity mention and its candidate. The performance of our approach achieves significant improvement compared to the best scores in the BB-norm shared task of BioNLP-OST19.

[1] https://sites.google.com/view/bb-2019/home

2 Approach

Our approach includes two principal steps: (1) **Candidate concept generation:** for a given biomedical entity mention, generating candidate concepts from the KB or ontology. (2) **Candidate concept ranking:** ranking those candidate concepts. Further details are provided in the following sections.

2.1 Candidate Concept Generation

We first pre-process all entity mentions and concept names in KB with abbreviation and tokenization resolution. The *Ab3p* tool[2] is utilized to identify abbreviations in documents and replace entity mentions in abbreviations with their corresponding full names. The *Snowball toolkit*[3] is used to tokenize all the entity mentions and concept names. Then, we implement two types of methods to generate the candidate concepts:

Similarity based method: We calculate the cosine similarity between vector representations of each concept name and the mention, and also the Jaccard similarity between them [5]. Then we choose the top n_1 concept names as a set C_1, which have cosine similarity greater than or equal to threshold t_1, top n_2 concept names as a set C_2, which have Jaccard similarity greater than or equal to threshold t_2. The final candidate set is composed of $C_1 \bigcup C_2$. In our experiments on training data, we set $t_1 = 0.7$, $t_2 = 0.1$, $n_1 = 3$, $n_2 = 7$.

Information Retrieval based method: It is inefficient to calculate the similarity between a given mention and each concept name in the KG when the number of concepts is very large. In this case, IR is a more efficient method with which to obtain similar concept names. We implement the IR system using Lucene[4]. First, we index all mentions in training data and concept names with their identities. Second, we retrieve the top-20 concept names for each mention, then the final candidate concept names are obtained from the retrieved results using the similarity based method described above.

2.2 Candidate Concept Ranking

We rank the candidate concepts by fine-tuning the pre-trained BioBERT model. Inspired by the work [6], for each entity mention m and one of its candidates c, we feed a sequence [CLS] m [SEP] c to BioBERT for the fine-tuning procedure, where [CLS] is the beginning of each sequence, and [SEP] is a special token used to separate m and c. V is supposed to be the final hidden state generated from BioBERT model, and d is the dimension of the hidden state of the model. $V_0 \in \mathbb{R}^d$ represents the output of the first token [CLS], V_m and V_c are the final hidden vectors for m and c respectively, $V_m = [V_i, ..., V_j]$ and $V_c = [V_l, ..., V_n]$.

[2] https://github.com/ncbi-nlp/Ab3P

[3] http://www.nltk.org/_modules/nltk/stem/snowball.html

[4] https://lucene.apache.org/

Then we get the representation of m ($V_m^{'}$) and the representation of c ($V_c^{'}$) using the following equations:

$$V_m^{'} = W[tanh(\frac{1}{j-i+1}\sum_{t=i}^{j} V_t)] + b \quad V_c^{'} = W[tanh(\frac{1}{n-l+1}\sum_{t=l}^{n} V_t)] + b \quad (1)$$

Where we use an average operation to the vectors, we then add an activation layer and a fully connected layer, there are $j-i+1$ words in m and $n-l+1$ words in c. For V_0, an activation layer and a fully connected layer are added:

$$V_0^{'} = W_0[tanh(V_0)] + b_0 \quad (2)$$

We then concatenate $V_0^{'}$, $V_m^{'}$, $V_c^{'}$ and add a fully connected layer to generate the final representation for a mention and one of its candidates:

$$\mathbf{r} = W_{con}[concat(V_0^{'}, V_m^{'}, V_c^{'})] + b_{con} \quad (3)$$

In Equations (1), (2), (3), W, $W_0 \in \mathbb{R}^{d\times d}$, $W_{con} \in \mathbb{R}^{3d}$, b, b_0, and b_{con} are bias vectors.

It is supposed that there are K candidates for each entity mention. We use r_k to represent the final vector output by our model for the k^{th} candidate name. To rank the K candidate names, we first define $R = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_K]$, and then compute the probability of the k^{th} candidate to be the normalized one as follows:

$$p(k|R) = sigmoid(W_r R + b_r) \quad (4)$$

Where $W_r \in \mathbb{R}^{3d}$, b_r is the bias. We use binary cross entropy as the loss function.

3 Experiments

Dataset: We use the Bacteria Biotope (BB) data to evaluate our approach. Three types of entities are involved: microorganism, habitat and phenotype. Microorganisms are normalized to taxa from the NCBI taxonomy[5], which contains 903,191 taxa plus synonyms. While habitat and phenotype entities are normalized to concepts from the OntoBiotope ontology[6] which includes 3,601 concepts plus synonyms. Table 1 shows the number of mentions, unique mentions and concepts for each entity type. In the candidate generation step, we use the IR based method to generate candidates for microorganisms, and the similarity based method to generate candidates for habitat and phenotype entities.

Metrics: Since the entity mentions in the data are given and every mention is normalized to a concept, we evaluate the performance of our biomedical concept normalization algorithm with precision, following the BB-norm task. The official on-line testing platform[7] is used to calculate scores on the test data.

[5] `ftp://ftp.ncbi.nih.gov/pub/taxonomy`

[6] `http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE`

[7] `http://bibliome.jouy.inra.fr/demo/BioNLP-OST-2019-Evaluation/index.html`

Table 1. Statistics for each entity type.

	Habitat	Phenotype	Microorganism
Entity mentions	3,506	1,102	2,487
Unique entity mentions	1,774	498	950
Concepts	440	141	491

Parameters Settings: For fine-tuning, the parameters are the same as those in the pre-trained BioBERT model. We set the learning rate to 5e-5, the batch size to 16, and the number of training epochs to 16. Early stopping is employed according to the precision of the validation set.

Experimental Results The performance of the method is displayed in table 2. Compared to the methods from official teams who participated in the BB-norm shared task, our method achieves significant improvement of, respectively +4, +6, and +4 points compared to the best scores for habitat, phenotype, and microorganism normalization, and +10 points compared to the best score for all types. We also discard the hidden vector output of the entity mention and candidate, and only use the hidden vector output of the special first token for ranking, the results show that combining the output of entity and candidate vectors further enriches the information and improves the accuracy.

Table 2. Comparison of our biomedical entity normalization approach with the results in the BioNLP-OST19 challenge. Best scores are in bold font.

	All Type	Habitat	Phenotype	Microorganism
Baseline	0.531	0.559	0.581	0.470
BOUN-ISIK-2	**0.679**	**0.687**	0.566	0.711
BLAIR GMU-2	0.678	0.615	0.646	**0.783**
PADIA BacReader-1	0.633	0.684	**0.758**	0.511
Our approach(without entity)	0.762	0.708	0.821	0.817
Our approach	**0.778**	**0.733**	**0.823**	**0.825**

4 Conclusion

In this paper, we develop an approach for biomedical entity normalization by fine-tuning the pre-trained model, we also leverage the embeddings of entity mentions and their candidates to enrich the information and improve the performance. We conduct experiments on the BB dataset provided by the challenge of BioNLP-OST and our results significantly outperform the state of-the-art methods. In the future, we will try to add the context information of a mention to improve the performance of this problem.

References

1. Jennifer DSouza, Vincent Ng. Sieve-Based Entity Linking for the Biomedical Domain. In: Proceedings of ACL-IJCNLP15. pp. 279–302. Beijing, China (2015)
2. Robert Leaman, Rezarta Islamaj Dogan, Zhiyong Lu. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics. vol.29, pp. 2909-291. (2013)
3. Haodi Li, Qingcai Chen, Buzhou Tang. et al. CNN-based ranking for biomedical entity normalization. BMC Bioinformatics. (2017)
4. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. (2019)
5. Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, et al. Medical Entity Linking using Triplet Network. pp.95-100. (2019)
6. Shanchan Wu, Yifan He. Enriching Pre-trained Language Model with Entity Information for Relation Classification. CoRR. (2019)