Bacteria and Biotope Entity Recognition Using A Dictionary-Enhanced Neural Network Model

Qiuyue Wang and Xiaofeng Meng
School of Information, Renmin University of China
Beijing 100872, China
{qiuyuew, xfmeng}@ruc.edu.cn

Abstract

Automatic recognition of biomedical entities in text is the crucial initial step in biomedical text mining. In this paper, we investigate employing modern neural network models for recognizing biomedical entities. To compensate for the small amount of training data in biomedical domain, we propose to integrate dictionaries into the neural model. Our experiments on BB3 data sets demonstrate that state-of-the-art neural network model is promising in recognizing biomedical entities even with very little training data. When integrated with dictionaries, its performance could be greatly improved, achieving the competitive performance compared with the best dictionary-based system on the entities with specific terminology, and much higher performance on the entities with more general terminology.

1 Introduction

In the microbial community, knowledge about habitats of bacteria is crucial for the study, e.g. metagenomics. To extract such information from the biomedical literature, the very first step is to accurately recognize bacteria and habitat entities in text. State-of-the-art systems mainly have taken two approaches: dictionary-based and feature-based.

Dictionary-based approach looks for all the possible names in one or more dictionaries (or ontologies, or databases, or gazetteers) of entities. The performance depends on the quality and comprehensiveness of the dictionaries built for each entity type, which require a lot of expert knowledge and maintenance costs. It is well suited for entities with closely defined vocabularies of specific names, such as species and diseases, but fails to accurately recognize entities with names consisting of more common words, e.g. habitat entities. TagIt (Cook et al., 2016) is a dictionary-based system participating BioNLP Shared Task 2016, which yielded the best performance in recognizing bacteria entities, however could not compete with other machine learning systems on recognizing habitat entities.

Feature-based machine learning systems are currently more widely used in biomedical entity recognition. When properly trained, a machine learning model can potentially recognize new entity names and new spelling variations of an entity name. Traditional machine learning approaches, are feature-rich supervised learning classifiers, requiring significant domain-specific feature engineering. Recently neural network models gain increasingly more research attention as they could automatically learn useful features from raw data. Compared with the work on NER in general domain (Lample et al., 2016, Chiu and Nichols, 2016, Ma and Hovy, 2016), there is little published work on employing modern neural network models for BioNER. It is probably due to the small sizes of human-annotated corpora in biomedical domain, which makes it very hard to train non-trivial neural network models.

In this paper, we investigate employing state-of-the-art neural network models to recognize biomedical entities. Our experiments on BB3 data sets show that even with very little training data, modern neural network model is promising in recognizing bacteria and habitat entities. To compensate for the shortage of annotated training data, we propose to utilize dictionaries or ontologies, which is abundant in biomedical domain, to enhance the neural models. The experiment results demonstrate that our dictionary-enhanced neural model yielded better performance than the currently best systems, especially on habitat entities.

2 Dictionary-Enhanced BiLSTM-CRF Model

Following the most state-of-the-art neural network models for general domain NER, we design a similar BiLSTM-CRF model as shown in Figure 1 for recognizing bacteria and habitats in text.

Proceedings of the BioNLP 2018 workshop, pages 147–150

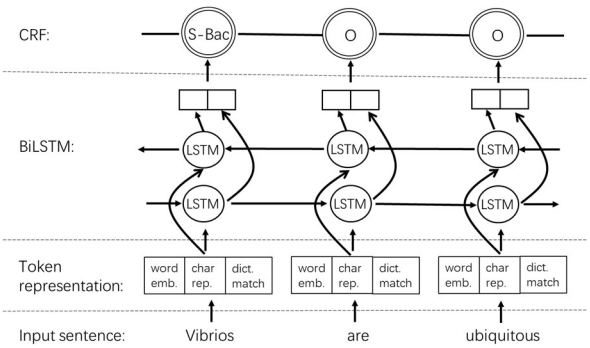

Figure 1: The BiLSTM-CRF model for entity recognition.

When receiving a sentence of tokens as input, e.g. "*Vibrios are ubiquitous*", the system first forms a representation for each token, which is the concatenation of its word embedding, character-based representation and dictionary-matching representation of the token.

Next, the vector representations of tokens are fed into a bidirectional LSTM. The hidden state for each token position in BiLSTM is the concatenation of the hidden states from the forward and backward LSTMs. As a result, it contains both the left and right context information useful to make prediction for this token.

Finally, a Conditional Random Field (CRF) layer, modeling the dependencies between successive labels, is added on top of the BiLSTM network to find the most likely sequence of labels as the final output.

2.1 Word Embedding

There are various word embedding techniques, e.g. word2vec, Glove and fasttext. They address different types of semantic similarities, and thus perform differently for different NLP tasks. We tested Glove and fasttext for our task and found that fasttext performed better. Thus, we used the fasttext method to train word embeddings. Word embedding dimension is set to 100.

We downloaded PubMed 2017 baseline, extracted all the titles and abstracts, segmented them into tokens using different strategies:

- using a segmentation model for general English text or a model specially trained on biomedical text.
- removing punctuations or not.
- converting all characters into lower case and all digits to "0" or not.

We compared the performance of all the above strategies in the experiments, and found that removing punctuations, lowercasing all characters and converting all digits to "0" did not result in better performed embeddings. So, we generated embeddings without removing punctuations and any other conversions.

2.2 Character-Based Representation

Although the word embeddings capture the semantic similarities between tokens, they ignore the character-level regularities of the token, like suffixes or prefixes, which are proven to be effective in NER tasks. We generate a character-based representation for each token using a LSTM model like that proposed in Lample et al., 2016. The dimensions of the character embedding and the hidden states of the BiLSTM are both set to be 25, so the dimension of the final character-level representation is 50.

2.3 Dictionary-Matching Representation

To train a non-trivial neural network model without overfitting it, a huge amount of annotated data are needed, which are much costlier to obtain in biomedical domain than in general domain since expert domain knowledge is required for annotating data. On the other hand, dictionaries, ontologies and databases are abundant in biomedical domain. We propose to make better use of such available knowledge in neural network models to compensate for the small sizes of annotated data.

In this paper, we incorporate dictionaries into the neural network model by adding a third part to the token representation: dictionary-matching representation. For each given dictionary, a dictionary matching feature is assigned to each input token. The matching feature indicates whether a word sequence formed by the token and its consecutive neighbors is in the dictionary. The maximal length of the word sequence is set to 6. When there are multiple overlapping matches, longer matches are preferred over shorter matches, and earlier matches in the sentence are preferred over later matches. The matching feature can take one of the five values: 'B', 'I', 'O', 'E', 'S', which means 'Begin', 'Inside', 'Outside', 'End' and 'Single' respectively, indicating the position of the token in the matched word sequence. Figure 2 shows an example sentence and the dictionary matching feature for each of its tokens. There are two types of entities to be recognized: bacteria and habitats, and two dictionaries are applied, one for each entity type.

	Vibrios	*are*	*ubiquitous*	*to*	*oceans*	*,*	*coastal*	*waters*	*,*	*and*	*estuaries*	*.*
Bacteria	S	O	O	O	O	O	O	O	O	O	O	O
Habitat	O	O	O	O	S	O	B	E	O	O	S	O

Figure 2: Dictionary matchings of an example sentence.

To generate the dictionary-matching representation for the token, we embed the matching feature for each dictionary into a 5-dimensional real-valued vector and then concatenate the vectors for all the dictionaries. As in Figure 2, the dictionary-matching representation of a token will be a 10-dimensional vector representing the matching features of this token in two dictionaries.

3 Experiments and Results

We implemented our models based on the open source code of NeuroNER[1] (Dernoncourt et al., 2017) and evaluated their performance using the dataset provided by the Bacteria Biotope task in the BioNLP Shared Task 2016 (BB3).

The BB3 task has no separate task for named entity recognition. It is jointly evaluated with downstream applications such as categorization or event extraction. Only in the BB3-cat+ner subtask, the official BB3 evaluation service additionally outputs the boundaries scoring about the system's ability to predict entity boundaries, in terms of SER (Slot Error Rate), Precision and Recall. For this reason, we primarily focus on the BB3-cat+ner subtask. We use the SER, Precision and Recall, output by the official BB3 evaluation service, as the evaluation metrics for our experiments. According to the official evaluation (Deléger et al., 2016), TagIt system achieved the best performance on detecting bacteria boundaries (SER: 0.236, recall: 0.772, precision: 0.954), while LIMSI system worked best on habitat entities (SER: 0.597, recall: 0.504, precision: 0.728). Bacteria are easier to recognize than habitats because bacteria names are mainly specific terms from a closely defined vocabulary, i.e. NCBI Taxonomy, with little variations, while habitat names usually consist of common English nouns and adjectives, e.g. "egg", "water", "fish" and expressed in various ways.

3.1 Dataset and Preprocessing

The dataset of the BB3-cat+ner subtask consists of 161 documents, split into training, development and test sets, which include 71, 36 and 54 documents and 1122, 698, 1022 entity occurrences respectively.

Entities occurring in the training or development documents are annotated in BRAT format. We preprocessed the data by first segmenting all the text into sentences of tokens using spaCy[2], and then tagging each token with a label in BIOES labelling scheme. For example, "B-Bacteria" means the token is the beginning word of a bacteria entity mention, and "S-Habitat" means the token is by itself the mention for a habitat entity.

3.2 Word Embeddings

For segmenting text to train word embeddings, we could use a segmentation model for general English text, or alternatively a model specifically trained on biomedical text. For the general model, we used spaCy, and for the specific model, we applied OpenNLP with its specially trained model on the GENIA corpus.

As shown by the first two lines in Table 1, using a specific model trained on domain text gained higher precision while lower recall than using a general English model. It also shows that the state-of-the-art BiLSTM-CRF model is a promising approach for recognizing biomedical entities, even with very little training data like in BB3 task.

3.3 Integration of Dictionaries

In general, performance of neural models could get far improved by using more training data. However, it is costly to collect a large amount of training data in biomedical domain. Recently, more and more research work focused on finding ways to compensate for the shortage of training data, e.g. using semi-supervised learning or multi-task learning techniques. In this paper, we exploited the way of integrating dictionaries or ontologies into the neural network model to improve performance. For detecting bacteria and habitats, we use the most recent comprehensive dictionaries[3] specially built for these two types of entities by TagIt. We tested two strategies of matching with dictionary entries: case-sensitive and case-

[1] http://neuroner.com/

[2] https://spacy.io/

[3] https://github.com/bitmask/BioNLP-BB3

Systems	Overall				Bacteria boundaries			Habitat boundaries		
	SER	**Recall**	**Prec.**	**F1**	**SER**	**Recall**	**Prec.**	**SER**	**Recall**	**Prec.**
spaCy	0.487	0.596	0.789	0.678	0.415	0.693	0.814	0.519	0.544	0.775
OpenNLP	0.493	0.549	0.830	0.661	0.376	0.656	0.891	0.558	0.490	0.785
spaCy (B+H)	0.435	0.624	0.828	0.712	0.324	0.701	0.919	0.503	0.578	0.768
spaCy (B+H, lower)	0.429	0.617	0.852	0.715	0.318	0.710	0.918	0.499	0.556	0.801
OpenNLP (B+H)	0.442	0.578	**0.876**	0.697	0.330	0.684	0.925	0.514	0.511	**0.835**
OpenNLP (B+H, lower)	**0.415**	0.617	0.867	**0.721**	0.301	0.707	**0.938**	0.483	0.563	0.816
TagIt (Cook et al., 2016)	-	-	-	-	0.236	0.772	0.954	0.599	0.476	0.675
LIMSI (Grouin, 2016)	-	-	-	-	0.277	0.751	0.903	0.597	0.504	0.728

Table 1: Experiment results.

insensitive matching. In Table 1, "B+H" represents for using bacteria and habitat dictionaries, and "lower" means case-insensitive matching with the dictionary.

From Table 1, we can have the following observations:

(1) By comparing the "B+H" lines with the first two lines, we can observe that integrating dictionaries into neural models can significantly improve the performance. For example, the overall SER is reduced by 12%-16%.

(2) By comparing the "B+H" lines with "B+H, lower" lines, we see that case-insensitive matching with dictionary is more effective than case-sensitive matching.

(3) Compared with the existing two best systems using traditional dictionary-based (TagIt) or feature-based (LIMSI) approaches, our best model "OpenNLP (B+H, lower)" can perform competitively on recognizing bacteria entities and much better on recognizing habitat entities.

4 Conclusions and Future Work

To the best of our knowledge, this is the first work of applying state-of-the-art neural network models in recognizing bacteria and biotope entities. The experiment results on BB3 task show that it is promising even with very small sized training data. Its performance can be much improved by integrating dictionaries, achieving competitive performance on bacteria entities and much better performance on habitat entities compared with the best traditional methods.

As for future work, we intend to (1) test our model on more types of biomedical entities; (2) investigate other ways of integrating dictionaries or ontologies with neural networks; (3) extend our model to deal with the embedded entities and discontinuous entities, which are special challenges for BioNER.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61532010), the National Key Research and Development Program of China (No. 2016YFB000603), and the Opening Project of State Key Laboratory of Digital Publishing Technology.

References

Jason P.C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *TACL*, vol. 4, pp. 357–370.

Helen V Cook, Evangelos Pafilis, Lars J. Jensen. 2016. A dictionary- and rule-based system for identification of bacteria and habitats in text. *BioNLP 2016*.

Louise Deléger, et al. 2016. Overview of the Bacteria Biotope task at BioNLP Shared Task 2016. *BioNLP 2016*.

Franck Dernoncourt, Ji Young Lee, Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. *EMNLP 2017*.

Cyril Grouin, 2016. Identification of Mentions and Relations between Bacteria and Biotope from PubMed Abstracts. *BioNLP 2016*.

Guillaume Lample, et al. 2016. Neural architectures for named entity recognition. *NAACL-HLT 2016*, pages 260–270.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *ACL 2016*, pages 1064–1074.