Recurrent neural networks with specialized word embedding for Chinese Clinical Named Entity Recognition

Zhenzhen Li, Qun Zhang, Yang Liu, Dawei Feng, and Zhen Huang

College of Computer, National University of Defense Technology, China
Changsha, Hunan 410073
lizhenzhen14@nudt.edu.cn,chriszhang0511@gmail.com,
{jokertail,davyfeng.c}@qq.com
maths_www@163.com

Abstract. To extract medical clinical related entity mention from patient clinical records is an essential step in clinical research. Recently, many researchers employ neural architecture to tackle the similar task of clinical concept extraction or drug name recognition from English clinical records, and have got prominent progress. However, most previous systems on Chinese Clinical Named Entity Recognition(CNER) have focused on a combination of text "feature engineering" and conventional machine learning algorithms. In this paper, we proposed a nerual network system based on bidirectional LSTMs and CRF for Chinese CNER. Also, we use health domain datasets to create richer, more specialized word embeddings, and combined with the external health domain lexicons, the performance is futher improved.

Keywords: Neural networks, Chinese Clinical Named Entity Recognition, specialized word embedding

1 Introduction

Patient clinical records contain longitudinal record of patient health, disease, conducted test and response to treatment, often useful for epidemiologic and clinical research. Thus extracting these information has been of immense value for both clinical practise and to improve quality of patient care provided, while reducing healthcare costs. The evaluation task of Clinical Named Entity Recognition (CNER) aims to identify medical clinical related entity mentions from Electronic Health Record narratives, and classify them into predefined categories, such as disease, symptoms, examination etc.

Traditional approaches to Named Entity Recognition(NER) relied on rule based systems or dictionaries (lexicons) using string comparision to identify entity mention of interest. Although these systems achieve high precision, they still suffer from low recall and are hard to scale. Many related research regards NER as a sequence labelling task. The applied methods on NER include Conditional

Random Fields (CRFs), Hidden Markov Models (HMMs), and they try to jointly infer the most likely label sequence for a given sentence.However, these methods rely heavily on hand-crafted features and task-specific resources, which is costly to develop. To overcome these limitations, neural networks are applied to this task [1] and achieved competitive performance. This paper employs bidirectional LSTM-CRF for automatic feature learning thus avoiding time-consuming feature engineering.

As to Chinese NER, there are more complicated properties in Chinese, for example, the lack of word boundary, the complex composition forms, the uncertain length and so on [2]. To obtain the representations that embedded more precise semantics, the sentences are segmented to words or phrases by NLP tools, and each word or phrase can be represented by a numerical vector (the embedding). Since the task of CNER is limited to the specific domain, the CNER systems should also focus on text with specific dictionaries and topics, together with dedicated sets of named entities. The NER system [3] has provided evidence that retrained embeddings on a domain-specific dataset can help learn vector representations for domain-specific words and increase the classification accuracy. Therefore, this paper crawls Chinese health domain corpus to create richer, more specialized word embeddings.

In this paper, we develop a neural network architecture for Chinese clinical named entity recognition with more complex and specialized word embeddings. Moreover, we use the external clinical lexicons to optimize the extracted entity mentions, which proves to be effective. External clinical lexicons are also used to label the unlabeled dataset, so we get a noisy considerable training dataset, which improves the neural architecture's performance.

2 Model

Our neural network is inspired by the work of Lamplea et al.(2016) [1], where feature vectors are computed by lookup tables and concatenated together, and then fed into a BiLSTM neural network. Instead of using the general-purpose, pre-trained word embeddings, we retrained word embeddings on the health related and clinical dataset.

2.1 Input Word Embeddings

specialized Word Embeddings A word embedding maps a word to a numerical vector in a vector space, where semantically-similar words are expected to be assigned similar vectors. To perform this mapping, we use the gensim tool and choose the CBOW(Continuous Bag-Of-Words) algorithm to train a word2vec model.

The datasets used to train the embeddings consist of two parts. One is crawled from several Chinese health related or clinical websites, such as 39

health[1], healthcare and learning[2] , medical encyclopedia[3] and so on. The crawled corpora contain the descriptions about diseases, the symptoms, the therapeutic methods and the doctor's responses to the messages, which amount to over one million sentences. The other corpus is published by the CNER Evaluation, including the labeled data and unlabeled data. All these corpora are preprocessed by removing unwanted characters, such as special characters, punctuation and stop words. Then the specialized word embeddings are trained on these datasets, which contain over 20 thousand unique words.

As a matter of fact, in health corpora it is common to find some technical and unusual words which are specific to the health domain. Therefore, such datasets can generate good embeddings in many cases. However, for this domain-specific task of CNER they still suffer from some lack of vocabulary. These out of vocabulary words are initialized with random assignments, and we also use character-level embeddings to complement its semantics.

Character-level embeddings Following Lample et al. [1] we also add character-level embeddings of the words and phrases. The Chinese character means the minimum unit of a segmented word or phrase, which can reflect part semantics of the phrase. A character lookup table initialized at random contains an embedding for every character. The character embeddings corresponding to every character in a phrase are given in direct and reverse order to a forward and a backward LSTM. This character-level representation is then concatenated with a word-level representation from a word lookup table.

2.2 Bidirectional LSTM-CRF Networks

We provide a brief introduction on the hybrid tagging architecture, which is based on LSTMs and CRFs. The architecture is similar to the ones presented by Lample et al. [1], which show that NER can be successfully resolved as a sequence labeling task.

tagging scheme The task of CNER is to assign a named entity label to every word in a sentence. A single named entity could span several tokens within a sentence. Sentences are usually represented in the IOB format(Inside, Outside, Beginning). In this paper, we use IOBES(Inside, Outside, Beginning, Ending, Singleton) tagging scheme. Using this scheme, more information about the following tag is considered.

BiLSTM-CRF The Long Short-Term Memory (LSTM) is designed to learn long-term dependencies in the sequences by incorporating a gated memory-cell. They do so using several gates that control the proportion of the input to give to

[1] http://www.39.net/
[2] http://club.xywy.com/
[3] http://www.a-hospital.com/

the memory cell, and the proportion from the previous state to forget [4]. In this paper, we use one popular LSTM variant, introduced by [5], is adding "peephole connections". This means that we let the gate layers look at the cell state. We also use coupled forget and input gates. The implementation is followed by these Equations:

$$i_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \tag{1}$$

$$\tilde{C}_t = \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \tag{2}$$

$$C_t = i_t \odot \tilde{C}_t + (1 - i_t) \odot C_{t-1} \tag{3}$$

$$o_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o) \tag{4}$$

$$\mathbf{h}_t = o_t \odot \tanh(C_t) \tag{5}$$

where σ is the element-wise sigmoid function, and $\odot$ is the element-wise product. In the bidirectional LSTM, for any given sentence, the network computes both a left,$\overleftarrow{h_t}$, and a right,$\overrightarrow{h_t}$, representations of the sentence context at every input, x_t . The final representation is created by concatenating them as $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$.

The output vectors of BiLSTM are fed to the CRF layer to jointly decode the best label sequence. Since there are strong dependencies across output labels(e.g., I-PER cannot fowllow B-ORG), CRF layer can be used to utilize the dependencies and decide the final output label jointly. The labels are typically predicted using a Viterbi-style algorithm which provides the optimal prediction for the measurement sequence as a whole. Combining the continuous output label of the same type, we get the candidate entity mentions of interest.

2.3 optimizing with external knowledge

It is undeniable that external knowledge such as lexicons or the knowledge base is crucial to NER systems, especially on specific domains. Thus, we construct lexicons from annotated data and online data, which are related to the five categories (Symptom, Examination, Disease, Treatment, Body) defined by the CCKS 2017 CNER Evaluation task. The lexicons are applied to rectify some incorrect extracted entity mentions by rules. For example, the entity name "结核病史" is extracted by our system automatically, which is combined by two tokens, "结核 病史", while the gold label of the same position is "结核病". This is due to the wrong word segmentation. After word segmentation, if "病史" is regarded as one token, the system can never automatically pick out "结核病" as an entity name. With the help of the lexicons, such errors can be solved by replacing the extracted entity mention with the matched entry in the lexicon. A match is successful when the extracted entity mention overlaps an entry in the lexicon with the same type and the entry also appeares in the original sentence.

3 Evaluation

Evaluation was performed on the CCKS-2017 CNER shared task dataset. We ran each experiment multiple times and remained the best hyper-parameters. We also applied a dropout mask to the final embedding layer just before the input to the BiLSTM network.

3.1 Datasets

The training set has 1198 electronic medical records, and the test set has 796 records. The five predefined categories are symptoms and signs, examination and inspection, disease and diagnosis, treatment, body parts, which are abbreviated as Symptom, Examination, Disease, Treatment, Body. Table 1 shows the quantity of entity mentions labeled in the datasets from 5 categories.

Dataset	Symptom	Examination	Disease	Treatment	Body
Train dataset	7831	9546	722	1048	10719
Test dataset	2311	3143	553	465	3021

Table 1. Size of each category in number of tokens (number of named entities)

As Table 2 shows with the raw training dataset, the performance of these two categories (Treatment, Disease) is poor, so we reasonably guess this is caused by the imbalance of training data. Therefore, we repeat all the sentences from the training dataset that contain the entity mentions in the Treatment category twice, and five times for the Disease category. After verifying the effectiveness of this method, we test our model on the revised training dataset.

For the dataset, we performed the following preprocessing:

- All sequences of digits 0-9 are replaced by a single "0".
- All sentences are segmented by the HanLP tool.

3.2 Evaluation Methodology

In this work, we employ the "strict" evaluation method, where both the entity class and its exact boundaries are expected to be correct. We report the performance of the model in terms of the F1-score.

3.3 Results and Analysis

Table 2 shows results with different training dataset. After adding more data of the two categories which have less entity mentions, the dataset tends to be more balanced considered these categories. The values of F1-score in the three categories, Disease, Treatment, Body, have obviously increased. Though slightly dropping down in Symptom and Examination, the F1-score of overall increases.

Dataset/F1-score	Symptom	Examination	Disease	Treatment	Body	Overall
raw training dataset	0.9094	0.9000	0.6728	0.6686	0.7393	0.8250
revised training dataset	0.8199	0.8734	0.7054	0.8607	0.8049	0.8301

Table 2. Results with different training dataset

Thus we can conclude that this method of revising the training dataset is effective to the task. Our architecture have several components that have different impact on the overall performance. Table 3 presents our comparison with different word embedding. Compared with random assignments, our model obtained better performance with specialized word embedding by +2.7%. With the help of the lexicons, the F1-score gained by 2.2%

Variant	F1
BiLSTM-CRF + random	0.8301
BiLSTM-CRF + specialized	0.8575
BiLSTM-CRF + specialized+Lexicons	0.8795

Table 3. Results with different methods

4 Conclusion

In this paper, we have set to investigate the effectiveness of the BiLSTM-CRF architecture with specialized word embedding for Chinese clinical named entity recognition, and compared them with a baseline neural network model. As input features, we have applied combinations of specialized word embedding with character-level embedding. And the flexible application of the lexicons also improves the performance further.

References

1. Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[J] In Proc. of NAACL-2016, San Diego, California, USA, June.
2. Duan H, Zheng Y. A study on features of the CRFs-based Chinese Named Entity Recognition[J]. International Journal of Advanced Intelligence, 2011, 3(2): 287-294.
3. Unanue I J, Borzeshi E Z, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition[J]. CoRR:abs/1706.09569, 2017.
4. S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.
5. Gers F A, Schmidhuber J. Recurrent nets that time and count[C]//Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on. IEEE, 2000, 3: 189-194.