DUTIR at the CCKS-2018 Task1: A Neural Network Ensemble Approach for Chinese Clinical Named Entity Recognition

Ling Luo, Nan Li, Shuaichi Li, Zhihao Yang* and Hongfei Lin

College of Computer Science and Technology, Dalian University of Technology, Dalian, China 116024
*Corresponding Author: yangzh@dlut.edu.cn

Abstract. As a fundamental task in information extraction, named entity recognition (NER) has received constant research attention over the recent years. The 2018 China conference on knowledge graph and semantic computing (CCKS) challenge sets up a task for clinical named entity recognition (CNER). In this task, we presented a neural network ensemble approach, which combines five individual neural network models (i.e., CNN-CRF, BiLSTM-CRF, BiLSTM-CNN-CRF, BiLSTM+CNN-CRF and Lattice LSTM). In this approach, the various features (i.e., stroke, word segmentation and dictionary features) are adopted. On the official test set, our best submission achieves the F-scores of 88.63% and 95.19% under the "strict" and "relaxed" criteria, respectively.

Keywords: Entity Recognition, Chinese Clinical Text, Neural Network, Ensemble.

1 Introduction

In recent years, the medical information processing has become a popular research focus as the generation of larger amount of electronic medical records and the potential requirements for medical information services and medical decision supports. As a fundamental task for medical information extraction, clinical named entity recognition (CNER) has received much attention and has been organized as a shared-task in many challenges [1-3]. To promote the performance of CNER on the Chinese clinical text, the 2018 China conference on knowledge graph and semantic computing (CCKS-2018) organized a CNER task to identify and extract the related medical clinical entities (i.e., anatomy, symptom, independent symptom, drug and operation) from Chinese clinical text.

In the previous work, the state-of-the-art CRF-based NER methods depend on effective feature engineering, i.e., the design of effective features using various NLP tools and knowledge resources, which is still a labor-intensive and skill-dependent task. Recently, deep learning has become prevalent in the machine learning research community. For the NER task, several similar neural network architectures [4-6] have been proposed and exhibit promising results. Compared with the traditional CRF-

based methods, the key advantage of these deep learning methods is that these layers of features are not designed by human engineers and, therefore, the least feature engineering is needed.

In this paper, we describe our CNER method for the CCKS-2018 CNER task. In our method, five individual neural network models (i.e., CNN-CRF, BiLSTM-CRF, BiLSTM-CNN-CRF, BiLSTM+CNN-CRF and Lattice LSTM) are used for CNER. Afterwards an ensemble model is built by combining these models' results with majority voting. Moreover, we explored the effect of additional features to further improve the performance.

2 Methods

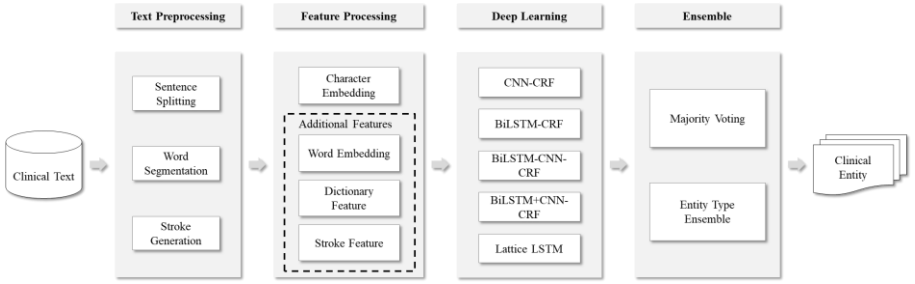

Fig. 1. The processing flow of our method

In this section, our approach for CNER is described. Fig. 1 shows the processing flow of our method. Firstly, some preprocessing steps including sentence splitting, word segmentation and stroke generation are performed. Secondly, a character embedding is learned with large amounts of unlabeled data by the cw2vec tool[1]. Moreover, the additional features (i.e., word embedding, dictionary feature and stroke feature) are introduced into the model. Then with the embeddings as input, five neural network models are trained by the annotated training set. Finally, the results from these models are combined by majority voting and entity type ensemble. The detailed description of our method is presented in the following sections.

2.1 Features

The character embedding is used as the basic features of our method since the character-based methods outperform word-based methods for Chinese NER [7, 8]. Moreover, to investigate the effects of other features (such as word embedding, dictionary feature and stroke feature), these features are added into the model as additional features. Details of each of features are provided as follows.

To achieve a high-quality character and word embeddings, we collected a total of 3,005 clinical texts from the CCKS-2017 challenge. Then these texts and all CCKS-2018 CNER training set (a total of 600 texts) were spilt into the words with the HanLP tool[2]. Afterwards, these data were used to train 100-dimensional character embedding and 50-dimensional word embedding by the cw2vec tool as pre-trained character and word embeddings.

Due to the complexity of the natural language and the specialty of the clinical

[1] https://github.com/bamtercelboo/cw2vec
[2] http://hanlp.linrunsoft.com/

medical domain, some linguistic and domain resource features can be employed to improve the performance of our model. We also explored the effect of dictionary and stroke features for our neural network models. For the dictionary feature, we used Sogou dictionary[3] to generate our drug dictionary feature. First, longest possible matches between the character sequences and dictionary entries were captured. Then, for each character in the match, the feature was encoded in BIOES tagging scheme. At last, a lookup table was used to output 50-dimensional dictionary embedding. For the stroke feature, we first obtained the stroke sequence of every character with HanDian[4]. Then a stroke lookup table containing a stroke embedding for every stroke was initialized randomly. Then the stroke embedding sequence of the character was fed into a convolutional layer. At last, an attention pooling layer is used to extract global features from the convolution layer as the stroke feature of the character.

2.2 NN-CRF Models

In this section we describe in details the five individual neural network models (i.e., CNN-CRF, BiLSTM-CRF, BiLSTM-CNN-CRF, BiLSTM+CNN-CRF and Lattice LSTM) used in our ensemble. These models have the similar neural network architecture, i.e., the neural network with a conditional random field layer (NN-CRF). The architecture of the NN-CRF model is illustrated in Fig. 2.

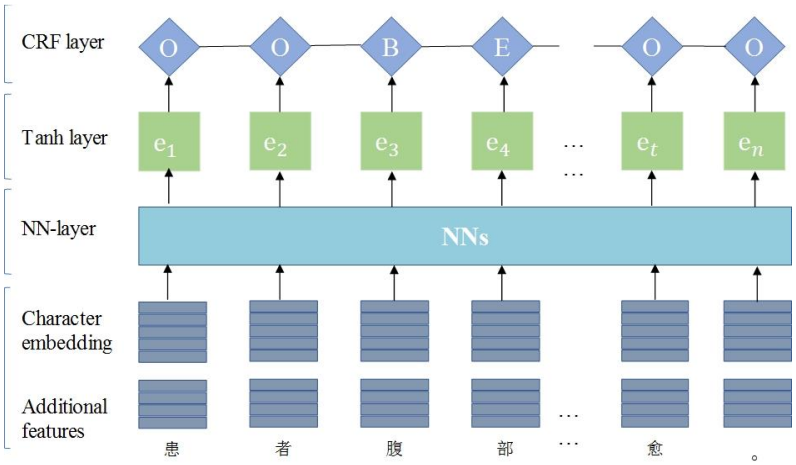

Fig. 2. The overall architecture of the NN-CRF model

- **CNN-CRF model.** In the convolutional neural network (CNN) with a CRF layer model, a convolution operation is applied to produce local features. Firstly, a sentence is represented as a sequence of embeddings. Next, the embeddings are given as input to a CNN layer. Then a tanh function on top of the CNN layer is used to learn higher features. Finally, the CRF layer is added after the tanh layer to predict the best label sequence path in all possible tag paths.

- **BiLSTM-CRF model.** We also employed a bidirectional long short-term memory with a CRF layer (BiLSTM-CRF) model for the CNER. Firstly, a sentence is

[3] https://pinyin.sogou.com/dict/detail/index/270
[4] http://www.zdic.net

represented as a sequence of embeddings. Next, the embeddings are given as input to a BiLSTM layer. In the BiLSTM layer, a forward LSTM computes a representation of the sequence from left to right, and another backward LSTM computes a representation of the same sequence in reverse. These two distinct networks use different parameters, and then the representation of a word is obtained by concatenating its left and right context representations. Then a tanh function on top of the BiLSTM layer is used to learn higher features. Finally, the CRF layer is added after the tanh layer to predict the best label sequence path in all possible tag paths.

- **BiLSTM-CNN-CRF model.** Different with the above models, the NN layer of this model is a BiLSTM-CNN layer. In the BiLSTM-CNN layer, a BiLSTM computes a representation of the sequence, which is then fed into a CNN layer to learn higher features.

- **BiLSTM+CNN-CRF model.** The model is similar with the BiLSTM-CNN-CRF model, but the BiLSTM-CNN layer is replaced with the BiLSTM+CNN layer. In the BiLSTM+CNN layer, the representation of the previous layer is fed into a BiLSTM layer and a CNN layer, and then their outputs are concatenated and fed into a tanh layer.

- **Lattice LSTM model.** Recently, Zhang and Yang proposed a lattice-structured LSTM model for Chinese NER [9]. In this model, the latent word information is integrated into character-based LSTM-CRF by representing lexicon words form the sentence using a lattice structure LSTM. This model explicitly leverages word and word sequence information, and does not suffer from segmentation errors.

2.3 Ensemble

As introduced before, five machine learning-based methods were deployed for the CNER independently. To take the advantages of different methods, we used a majority voting approach to combine all predicted entities. In addition, different combinations of the models were also investigated. Finally the models with the best performances for the five entity types (i.e., anatomy, symptom, independent symptom, drug and operation) on the development set were combined to output the final result.

3 Experiments

3.1 Dataset

In the CCKS-2018 CNER challenge, organizers provided a corpus including the training and test sets. The training set consists of 600 medical records annotated with five categories of entity, including anatomy, symptom, independent symptom, drug and operation. And the test set consists of 400 medical records. In our experiments, we randomly selected the 20% of the training set as the development set to tune the hyper-parameters. The statistics of entity on different categories are list in Table 1.

Table 1. Statistics of the entities of different categories

Dataset	Anatomy	Symptom	Independent	Drug	Operation
Training set	7,838	2,066	3,055	1,055	1,116
Development set	1,634	418	657	166	213

3.2 Evaluation

The evaluation metrics of this task include two criteria: 1) strict metrics which define a correct match as that the ground truth and extraction result share same mention, same boundaries and same entity type; 2) relaxed metrics which only consider the ground truth and the result have same entity type and overlap boundaries. All our evaluations were performed on the official test set using the evaluation tool of CCKS-2018 CNER challenge, which outputs micro-average precisions (Prec.), recalls (Rec.) and F-scores (F) via the strict metrics.

3.3 Experimental Results

In this task, we directly divided the sentences into Chinese characters, which can avoid the boundary error of entity caused by the word segmentation tools. The "BIOES" (i.e., B-begin, I-inside, E-end, S-single, O-outside) tags are used to represent the entity. Table 2 shows the performance of various models on our development set. Our basic model is BiLSTM-CRF which achieves an F-score of 90.88%. When stroke and dictionary features are added into the basic model, the model achieves performance improvement. The BiLSTM-CRF+Stoke+dic achieves an F-Score of 92.64%. It demonstrates that the features can help boost the performance of the model.

In addition, among the various models, BiLSTM-CRF+ALL and BiLSTM-CNN-CRF+ALL perform better than others for anatomy entities. BiLSTM-CRF+ALL and BiLSTM+CNN-CRF+ALL perform better for drug entities. To take the advantages of different methods, we used a majority voting approach to combine all predicted entities. The results show that the model ensemble achieves the highest F-score of 93.16%.

Table 2. Results of various models on our development set

Models	Anatomy	Symptom	Independent	Drug	Operation	Overall
BiLSTM-CRF	91.09	90.30	93.08	85.02	88.03	90.88
BiLSTM-CRF+Stroke	92.39	91.89	93.55	86.14	88.42	91.95
BiLSTM-CRF+Stroke+dic	92.66	92.88	93.92	91.89	88.63	92.64
CNN-CRF+ALL	92.58	**93.65**	94.90	91.94	87.62	92.83
BiLSTM-CRF+ALL	93.13	92.43	94.17	**92.77**	88.22	92.89
BiLSTM-CNN-CRF+ALL	93.07	91.72	93.94	91.19	89.20	92.70
BiLSTM+CNN-CRF+ALL	92.56	92.97	94.26	92.49	87.76	92.57
Lattice-LSTM	91.60	93.29	94.36	86.93	88.32	91.94
Ensemble	**93.21**	92.79	**94.98**	91.84	**89.67**	**93.16**

Note: "Stroke" denotes the stroke feature; "dic" denotes the dictionary feature; "All" denotes all additional features.

Table 3 lists the results on the official test set. Our best submission achieves the F-scores of 88.63% and 95.19% under the "strict" and "relaxed" criteria, respectively. We analyzed the results and found that the anatomy and operation have poor performance. Therefore, how to recognize the anatomy and operation entities more accurately will be the main focus in our future work.

Table 3. Results of our best submission on the official test set

Types	Strict (%)			Relaxed (%)		
	Prec.	Rec.	F	Prec.	Rec.	F
Overall	88.89	88.37	88.63	95.47	94.92	95.19
Anatomy	87.70	87.49	87.59	95.98	95.75	95.86
Symptom	92.73	88.89	90.77	94.77	90.85	92.77
Independent	91.52	91.93	91.72	94.67	95.10	94.88
Drug	92.69	90.41	91.53	95.21	92.87	94.92
Operation	85.62	86.67	86.41	93.68	94.83	94.25

4 Conclusion

In this paper, we present a neural network ensemble approach to automatically recognize clinical entities from Chinese clinical texts. In this approach, five different neural network models are explored, and the ensemble can achieve better performance. In addition, the effect of additional features for these models in the CNER task is also explored. The experimental results show that the additional features are effective to improve the performance of our system. At last, our best submission achieves the F-scores of 88.63% and 95.19% under the "strict" and "relaxed" criteria on the official test set, respectively. We will focus on the more effective extraction of anatomy and operation entities in the future work.

5 References

1. Uzuner Ö, Solti I, Cadag E: Extracting medication information from clinical text. *Journal of the American Medical Informatics Association* 2010, 17(5):514-518.

2. Sun W, Rumshisky A, Uzuner O: Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* 2013, 20(5):806-813.

3. Bethard S, Savova G, Chen W-T, Derczynski L, Pustejovsky J, Verhagen M: Semeval-2016 task 12: Clinical tempeval. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016): 2016*. 1052-1062.

4. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P: Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 2011, 12(Aug):2493-2537.

5. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C: Neural architectures for named entity recognition. *arXiv preprint arXiv:160301360* 2016.

6. Ma X, Hovy E: End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:160301354* 2016.

7. He J, Wang H: Chinese named entity recognition and word segmentation based on character. In: *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing: 2008*.

8. Li H, Hagiwara M, Li Q, Ji H: Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese. In: *LREC: 2014*. 2532-2536.

9. Zhang Y, Yang J: Chinese NER Using Lattice LSTM. *arXiv preprint arXiv:180502023* 2018.