Exploring N-gram Character Presentation in Bidirectional RNN-CRF for Chinese Clinical Named Entity Recognition

En Ouyang[1] , Yuxi Li[2], Ling Jin[1], Zuofeng Li[3] and Xiaoyan Zhang[1]

[1]Tongji University, Shanghai, China {1532751, 1632707, xyzhang}@tongji.edu.cn
[2]Peking University First Hospital, Beijing, China {liyuxi}@pku.edu.cn
[3]Philips Research China–Healthcare, Shanghai, China {zuo.feng.li}@philips.com

Abstract. Clinical named entity recognition (CNER) that identifies boundaries and types of medical entities, is a fundamental and crucial task in clinical natural language processing. Recent years have witnessed considerable progress in deep learning based algorithms, such as RNN, CNN and their integrated methods, which show the effectiveness in CNER. In this work, we propose a deep learning model for CNER that adopts bidirectional RNN-CRF architecture using concatenated n-gram character representation to capture rich context information. Second, we incorporate word segmentation results, part-of-speech (POS) tagging and medical vocabulary as features into our model. Further, the final output is delivered by the comparison between the separated models and the overall model. The proposed framework has been evaluated in CCKS2017 task2 dataset, achieving 90.10 F1-score for CNER.

Keywords: Clinical named entity recognition · Recurrent neural network · Conditional random field · N-gram

1 Introduction

Electronic medical record (EMR) systems have been adopted in most of hospitals in China and led to a huge amount of unstructured EMRs containing a wealth of important patient information[1]. As a consequence, access to these EMRs together with the machine learning techniques provides promising directions in clinical research and practice. Clinical named entity recognition (CNER), one of the most basic but crucial clinical natural language processing (NLP) tasks, is to identify the boundaries and types of entities.

CNER works on clinical records written in English have been studied for many years, bringing many existing CNER system applying various methods, such as MetaMap[2], a biomedical NLP system based on National Library of Medicine, along with cTAKES[3], an clinical NLP system developed based on the Unstructured Information Management Architecture (UIMA). The systems mentioned before both are based on dictionaries and rule-based methods, however those method may cause low recall and be definitely not flexible to various medical entities. Hence, many state-of-the-art CNER systems implement supervised machine learning (ML) algorithms with comprehensive handcraft features[4], [5]. Although ML methods with multiple features can attain remarkable performance, features construction can be a time-coursing process, which requires plenty of domain knowledge.

Addition to English CNER works, there is a growing interest in Chinese CNER studies. Conventional machine learning methods were used in Chinese CNER tasks, for instance, Lei et al. [6] compared different ML algorithms and various types of features for NER in Chinese admission notes and discharge summaries. With development of deep learning, typical deep learning models such as recurrent neural network (RNN) and conventional neural network (CNN) have achieved remarkable results in many natural language (NLP) tasks, including word segmentation[7], [8], POS tagging[7], sematic analysis[9] and text classification[10]. Wu et al. [11] proposed a deep neural network (DNN) method in Chinese CNER and compared it with the traditional CRF-based NER system at the minimal feature setting. Because the DNN architecture is relatively basic, the performance is not better than the CRF-based CNER system optimized using manual feature engineering[6].

In this work, a Bi-RNN-CRF architecture is applied in clinical named entity recognition. Character vectors represented as concatenated n-gram embeddings are feed into the neural network to leverage the context information. Furthermore, sematic features and medical vocabulary information are also incorporated and tested. Meanwhile, because of the difference between EMR categories, we train a model for each type of EMRs to select boosted model for final results. Then we also conduct post-processing based on medical knowledge to get final results.

The main contributions of this work are as follow:

1. We apply Bi-RNN-CRF model for medical entity recognition in electronic medical records (EMRs).

2. We adopt an approach for representation of character from Yan Shao's work[7] using n-gram information.

3. Additional improvements are obtained by using sematic and vocabulary features.

4. We integrate the results generated from models trained with four EMR groups respectively with the results generated from the model trained using all EMR data to get better prediction.

The paper is organized as follows. Section 2 describes the brief statistic of datasets, the basic idea and the details of our model based on Bi-RNN-CRF architecture. Next, we introduce the implementation and experiment results in Section 3. Finally, we summarize the key conclusions of our work in Section 4.

2 Methodology

2.1 Datasets

The organizers of CCKS Task2 provide 4 types of EMRs include: 一般项目(general items), 病史特点(medical history), 诊疗经过(diagnosis & treatment), 出院情况(discharge summary), in which there are 5 types of clinical entities to be recognized, include 身体部位(body), 症状和体征(symptom), 疾病和诊断(disease), 检查和检验(exam) and 治疗(treatment). **Table 1** shows the brief statistic of the training (in bold) and test (in standard) datasets in the number of EMRs and the number of entities in each EMR category. In addition to the labeled training and test datasets, 2205 unlabeled EMRs of each type were offered by the organizers. As listed in the **Table 1**, the entity ratio in EMR groups is different, for instance, there is a very few treatment and disease entities in Discharge summary, while the most entity in Diagnosis & treatment is treatment.

Table 1. Brief statistic of the training and test datasets.

	body	symptom	treatment	disease	exam
General items(**300**/100)	**181**/67	**558**/200	**0**/2	**74**/10	**1**/1
Medical history(**300**/100)	**6373**/1771	**4608**/1364	**138**/115	**570**/368	**5902**/1912
Diagnosis & treatment(**299**/99)	**875**/310	**547**/95	**902**/347	**74**/175	**794**/358
Discharge summary(**299**/99)	**3290**/873	**2118**/652	**8**/1	**4**/0	**2849**/872

2.2 Evaluation Measures

The evaluation metrics of this task include: strict metrics which define a correct match as that the ground truth and extraction result share same mention, same boundaries and same entity type, while relaxed metrics which define a correct match as that the ground truth and extraction result share same entity type and overlap boundaries. In this work, we evaluate the performance on f1-score of 5 entity categories and overall f1-score via the strict metrics.

2.3 Neural Network Architecture

Our basic model is an adaptation of Bi-RNN-CRF. As shown in **Fig. 1**, the clinical description texts are represented as vectors and fed into bidirectional RNN layers. The details of characters representation are described in following sections. The basic recurrent unit we used in the recurrent layers is LSTM to learn the long-term dependencies. Dropout is applied in text representation and the outputs of bidirectional recurrent layers. The output vectors of Bi-RNN are concatenated and passed to the CRF layer to jointly decode the best label sequence. Then, all entities along with their types and indexes in raw medical text will be retrieved.

Furthermore, the proportion of each entity type is different among four EMR categories, we propose a method to take advantage of this finding. As illustrated in **Fig. 2**, we train models for each type of EMRs respectively and a model for all EMRs (Model_ALL). If the validation F1-score of any model trained using just one EMR group is higher than the F1-score of Model_ALL, the model would be selected to generate results of this specific EMR group. Subsequently, those results will replace the results of that EMR category generated using Model_ALL in final output.

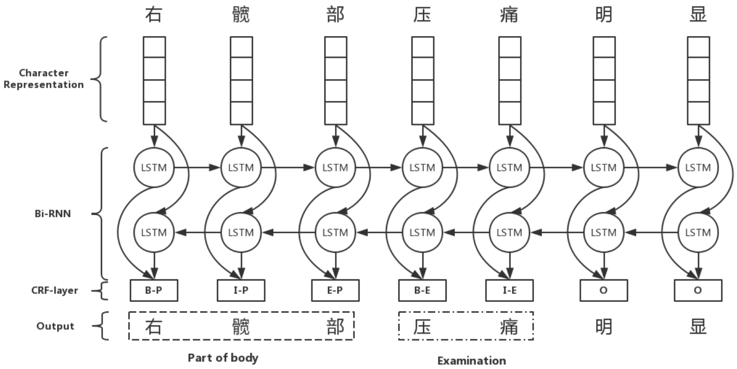

Fig. 1. The Bi-RNN-CRF model for medical entity recognition.

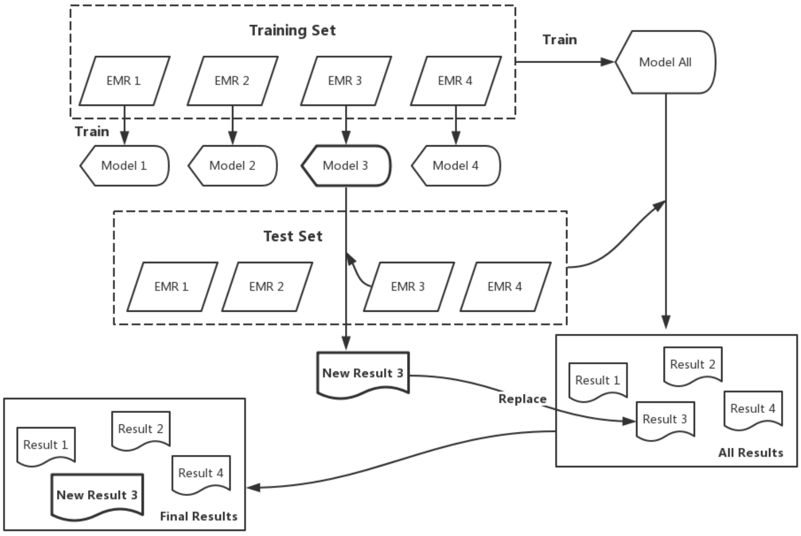

Fig. 2. The combination of Model_ALL and each type Model

2.4 Tagging Scheme

We utilize the following tags of entity boundaries: B, I, E, and S which respectively represent a character at the beginning, inside, end of an entity or as a single character entity. Then we combine the tags of entity type with the boundary tags as our training tags. For instance, the tags of "右髋部" are "B-body I-body E-body" and the tag of "脑" is "S-body".

2.5 Character Representation

To encode the local information for characters, we employ an incrementally concatenated n-gram representation. Word embeddings is often a key method to conduct word representation, so in this study we pre-train the word embeddings in all datasets using word2vec method, and calculate the occurrence frequency of every characters. For every character w_i in sentence, the single vector representation of w_i is $V_{i,i}$. If the frequency of w_i is higher than *min_freq* (minimum character occurrence frequency) cutoff, $V_{i,i}$ would be the word embedding of w_i. If not, $V_{i,i}$ would be the word embedding of <*UNK*> (the average of all word embeddings). $V_{i-1,i}$ is a concatenation of $V_{i-1,i-1}$ and $V_{i,i}$, $V_{i-1,i+1}$ is a concatenation of $V_{i-1,i-1}$, $V_{i,i}$,and $V_{i+1,i+1}$. The character representation V of w_i is a concatenation of $V_{i,i}$, $V_{i-1,i}$ and $V_{i-1,i+1}$. If the character is beginning or end of texts, it will be padded to bi-gram and tri-gram mentioned before. As demonstrated in **Fig. 3**, the vector representation of character 压 in the given context is the concatenation of the context-free vector representation $V_{i,i}$ of 压 itself, $V_{i-1,i}$ of the bi-gram 部压 along with $V_{i-1,i+1}$ of the tri-gram 部压痛.

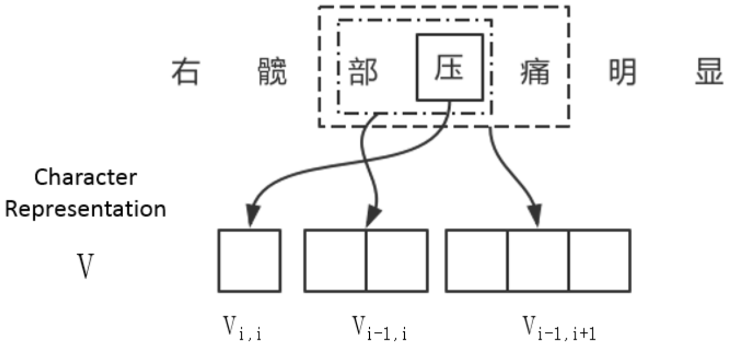

Fig. 3. Vector representations of the Chinese characters as concatenated n-gram vectors.

2.6 Additional Features

First, word segmentation information are used as extra features, we employed B, I, E, S tagging method to indicate the word boundaries. In vector representation, we used one-hot vector, for instance, if the character is beginning of a word, the vector should be [1, 0, 0, 0]. Second, we find that most entities begin with a noun, if a character is beginning of a noun, vector representation should be [1], if not, the vector should be [0]. Third, we collect a medical entities vocabulary containing Disease, Drug, Part of Body, the sources of entities are standard and freely accessed, including ICD-10, ICD-9-CM and other system. If a character is a character of those entities found in EMR texts, the vector representation should be [1], if not, the vector should be [0]. All feature vectors would be concatenated to a vector with the dimension of 6, which will be concatenated with the n-gram vector mentioned before to generate the vector representation of each character.

3 Implementation and Experiments

3.1 Implementation and Parameters

Our neural networks are implemented using Python3.5.2 and the TensorFlow 1.1.0-rc0 library, Chinese word segmentation and POS tagging are implemented using Jieba 0.38. We cut the sentences to segments as long as possible, but the length of which should be shorter than the defined maximum segment length, using Chinese sematic segmentation punctuations, such as comma and period. The segments in last step are padded to the defined maximum length. **Table 2** shows the adopted hyper-parameters. Just one set of parameters were used for each type model and Model_ALL. We found that the model trained using 出院情况 (discharge summary) EMRs outperformed the Model_ALL, so the results of 出院情况 (discharge summary) EMRs was generated from this model.

3.2 Baselines

We compare our model with three baselines:

1. *Bi-RNN-CRF*: A sample model using Bi-RNN-CRF architecture based on word embedding character representation.

2. *Bi-RNN-CRF_N*: A model using Bi-RNN-CRF architecture based on concatenated n-gram character representation.
3. *Bi-RNN-CRF_N_F*: A model similar to the proposed model in Section 2, except the results combination in final output.

Table 2. Hyper-parameters.

Parameter	Value
Character embedding size	100
Dropout rate	0.5
Initial learning rate	0.001
Batch size	30
Maximum training epochs	200
Maximum segment length	70
Minimum occurrence frequency(*min_freq*)	4

3.3 Experimental Results

As listed in **Table 3**, we compare the performance of our method with the baselines described in Section 3.2 in terms of 5 entity categories and overall f1-score using strict metrics.

We first observe that our model outperforms all the baselines in overall f1-score. In addition, n-gram character representation, sematic features and results combination led to the performance improvement of basic Bi-RNN-CRF model. We also observe that the f1-score of symptom and exam is better than the f1-score of other entities. Considering the imbalanced entity composition in training dataset, we can conclude that insufficient distribution of an entity category in training datasets can decrease its recognition property.

Meanwhile, we find that Discharge summary EMRs contain enough training data for body, symptom and exam, along with almost no mention of disease and treatment, in this way, the number of entity category can be reduced to 3. So that the model from Discharge summary can get a better performance than the Model_ALL in Discharge summary. When the scale of data is larger, training models for each type of EMRs and a model for all EMRs could face the problem of efficiency. In order to solve this problem, we should selectively train models for EMR categories, which satisfy the characteristics mentioned before, enough training data and less entity types.

Furthermore, the consistency of annotation strategy is crucial in training NER model, so we investigate the datasets annotation strategy. We found several inconsistency in all the datasets, firstly, there are entities without annotated in one or more than one EMR groups while they are exactly annotated in other type of EMRs. For instance, the entity “震颤” was not annotated in sentence like “未触及震颤” in Discharge summary, however all of them were annotated in Medical history. The same difference occurred in “叩心界”, “化痰”, “挫伤”, “血糖” and “尿痛”. Secondly, the annotation strategy of complex entities may be inconsistent, for instance, “胸廓对称” was annotated as one entity in Medical history, whereas it could be separated into two entities “胸廓” and “对称” in other type of EMRs. In order to achieve higher performance, a new annotation strategy with higher consistency would be applied.

Table 3. Comparison between our model and baselines.

Model	body	symptom	disease	exam	treatment	overall
Bi-RNN-CRF	0.8332	0.9473	**0.7622**	0.9274	0.7337	0.8843
Bi-RNN-CRF_N	0.8352	0.9457	0.7470	**0.9328**	0.7497	0.8864
Bi-RNN-CRF_N_F	**0.8377**	0.9481	0.7610	0.9299	**0.7551**	0.8877
Bi-RNN-CRF_N_F(combined)	0.8361	**0.9507**	0.7610	0.9319	**0.7551**	**0.8885***

*the public score is 0.9010

4 Conclusions

In this study, we proposed a clinical named entity recognition system based on Bi-RNN-CRF architecture. In consideration of the context information of medical texts and difference in four EMR categories, we apply a concatenated n-gram character representation method and combine the results from models trained using each EMR group. The experiment results showed that our approach outperforms the baselines. Notably, we find the character representation introduced context information would benefit the model. Moreover, the unbalanced distribution of entity type would influence the performance, and introducing EMRs type information would improve the overall performance. For future work, we aim to investigate more effective character representation and feature engineering method to speed up the training process. On the other hand, we will extend the usage of EMRs type information in other processes, such as word embeddings generation. Interestingly, semi-supervised and active learning methodology will be further explored to improve the CNER effectiveness via the incorporation of unlabeled data sets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (81573023 and 31601075) and National Key R&D Program of China (Grant No.2016YFC1303200). Meanwhile, we would like to thank Liang Tao, Yiyi Hu and Xiaomin Li for their insightful comments and guidance.

References

[1] China Hospital Information Management Association (CHIMA), “Investigation of hospital informatization in China (2014-2015).” Jun. 2015.

[2] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: historical perspective and recent advances,” *J. Am. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, May 2010.

[3] G. K. Savova *et al.*, “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 17, no. 5, pp. 507–513, 2010.

[4] J. D. Osborne, M. Wyatt, A. O. Westfall, J. Willig, S. Bethard, and G. Gordon, “Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning,” *J. Am. Med. Inform. Assoc.*, p. ocw006, Mar. 2016.

[5] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, “Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features,” *BMC Med. Inform. Decis. Mak.*, vol. 13, no. 1, p. S1, Apr. 2013.

[6] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and H. Xu, “A comprehensive study of named entity recognition in Chinese clinical text,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 21, no. 5, pp. 808–814, Sep. 2014.

[7] Y. Shao, C. Hardmeier, J. Tiedemann, and J. Nivre, “Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF,” *ArXiv Prepr. ArXiv170401314*, 2017.

[8] Y. Yao and Z. Huang, “Bi-directional LSTM Recurrent Neural Network for ChineseWord Segmentation.” .

[9] X. Wang, W. Jiang, and Z. Luo, “Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts.,” in *COLING*, 2016, pp. 2428–2437.

[10] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent Convolutional Neural Networks for Text Classification.,” in *AAAI*, 2015, vol. 333, pp. 2267–2273.

[11] Y. Wu, M. Jiang, J. Lei, and H. Xu, “Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network,” *Stud. Health Technol. Inform.*, vol. 216, pp. 624–628, 2015.