HITSZ_CNER: A hybrid system for entity recognition from Chinese clinical text

Jianglu Hu, Xue Shi, Zengjian Liu, Xiaolong Wang, Qingcai Chen, Buzhou Tang*

Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China, 518055
{hujianglu.hit, shixue.hit, liuzengjian.hit, qingcai.chen, tangbuzhou}@gmail.com, wangxl@insun.hit.edu.cn

Abstract. With rapid development of electronic medical records, more and more attention has been attracted to reuse these data for research and commercial. As the entity recognition is one of the most primary task for medical information extraction, the 2017 China conference on knowledge graph and semantic computing (CCKS) challenge sets up a track for clinical named entity recognition (CNER). The organizers provide 400 annotated Chinese medical records for this track, 300 out of them are used as a training set and 100 as a test set. Other 2,605 raw medical records are released as an unlabeled set. In this study, we develop a hybrid system based on rule, CRF (conditional random fields) and RNN (recurrent neural network) methods for the CNER task. Experiments on the official test set show that our system achieves the F1-scores of 91.08% and 94.26% under the "strict" and "relaxed" criteria respectively, ranking first in the 2017 CCKS CNER challenge. By applying a self-training method with unlabeled data, the F1-scores of all machine learning-based methods are improved by about 1.0% under "strict" criterion. The future work of us will focus on the more effective extraction of body, disease and treatment entities.

Keywords: Entity Recognition, Chinese Clinical Text, Recurrent Neural Network, Conditional Random Fields, Hybrid Method.

1 Introduction

In recent years, the medical information processing has become a popular researching focus as the generation of larger amount of electronic medical records and the potential requirements for medical information services and medical decision supports. Clinical entity recognition, one of the most primary clinical text processing task, has been organized as a shared-task in many challenges, such as the i2b2 (the center for informatics for integrating biology & the beside) 2009[1], i2b2 2010[2], i2b2 2012[3], SHEL (ShARe/CLEF eHealth Evaluation Lab) 2013[4], SemEval (Semantic Evaluation) 2014[5], SemEval 2015[6], SemEval 2016[7], etc. These challenges not only accelerate the research on entity recognition, but also annotate several valuable corpo-

* Corresponding author, Email: tangbuzhou@gmail.com

ra for clinical entity recognition. However, no one of them was organized on the Chinese clinical text. For this purpose, the 2017 CCKS (the China conference on knowledge graph and semantic computing) challenge sets up a clinical named entity recognition track (CNER) to identify entities from Chinese clinical text, in which five categories of entity are defined: Body, Disease, Symptom, Test and Treatment.

The early clinical entity recognition systems are mainly based on dictionary-matching and rules, such as MedLEE[8], etc. The machine learning-based clinical entity recognition systems have been developed since the last few years, especially after above several clinical entity recognition challenges have been organized. The main machine learning algorithms used for entity recognition include: hidden markov model (HMM), conditional random field (CRF) and structured support vector machine (SSVM), etc. In recent years, the recurrent neural network (RNN) has been widely used for clinical entity recognition, and achieves the state-of-the-art performances on i2b2 2010, i2b2 2012 and i2b2 2014 corpora[9].

In this study, we participate in the 2017 CCKS CNER challenge and develop a hybrid system for the Chinese clinical entity recognition, which is based on four individual methods (rule, CRF, RNN and RNN with features) and a vote-based approach. Besides, we also apply a self-training method with a large unlabeled dataset to improve the performance of our system.

2 Methods

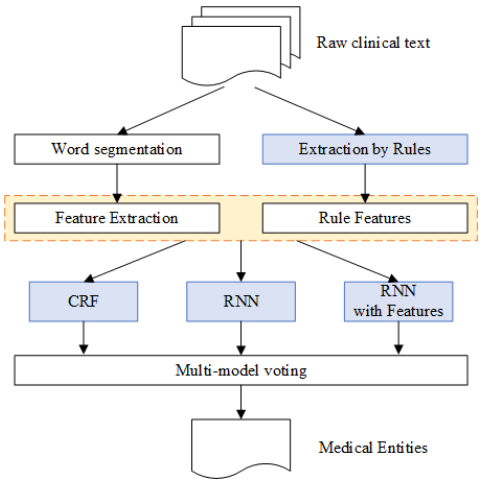

Fig. 1. The overview architecture of our system for the clinical entity recognition.

Figure 1 shows the overview architecture of our system for the entity recognition from Chinese clinical text, which contains four individual methods: rule-based, CRF-based, RNN-based and RNN with features methods. Firstly, we deploy these methods on Chinese clinical text respectively, the results of rule-based method are used as the features in other three methods. Then a vote-based method is used to combine all predicted entities by them. The detailed description of our system is presented below.

2.1 Rule-based Method

Since the prior-knowledge plays an important role in the entity recognition, especially for the clinical text, we construct several dictionaries for each type of entity referring to the training set and some open websites (e.g. “Baidu baike”, “Xunyiwenyao”, etc.), such as: body location, disease, symptom, examine, surgery, medicine, etc.

Assisting by these dictionaries, we build lots of rules to recognize the common patterns of entities, for example, in the phrase of “右侧小脑” (“right epencephalon”), “小脑” (“epencephalon”) can be identified as “Body” by dictionary-matching, the “右侧” (“right”) will be extended by our rules. Besides, in “…有心脏病病史…” (“...has history of heart attack…”), we can extract the “心脏病” (“heart attack”) as “Disease” according to the pattern “…有…病史…” (“…has history of …”).

2.2 CRF-based Method

As the CRF algorithm has been widely used for sequence labeling tasks, we also develop a CRF-based method for the clinical entity recognition using CRF++ as the implementation of CRF. The features used in this method include: n-gram, radical feature, spelling feature, word segmentation, part-of-speech, section head, dictionary feature, relation feature, distributed representation of word, rule feature, etc.

2.3 RNN-based Method

In this paper, we also employed a bidirectional LSTM (BI-LSTM), long short-term memory - a variant of RNN, method for the entity recognition from Chinese clinical text, which consists of three main layers: 1) input layer, generates the representation of each word in a sentence; 2) LSTM layer, takes the word representation sequence as input and generates a new one that captures the context information of the words. It contains both forward and backward LSTM networks; 3) output layer, learns the dependencies between successive labels by a transition matrix, and predicts a best label sequence according to the output of LSTM layer.

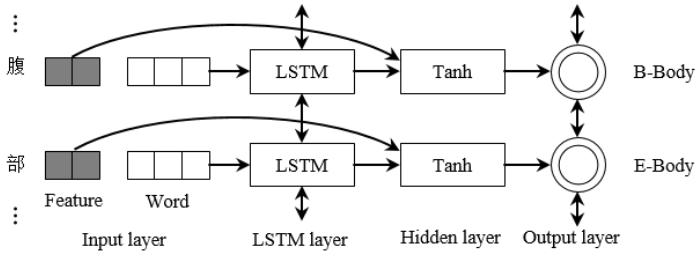

Fig. 2. The overview architecture of BI-LSTM-FEA network.

To utilized the hand-crafted features constructed in above rule and CRF based models, we extend above neural network (BI-LSTM) by adding a hidden layer (fully connected layer, FC) after the LSTM layer to concatenate the feature representations, which is represented as BI-LSTM-FEA, as shown in Figure 2.

2.4 Voting and Self-training

As introduced before, three machine learning-based methods are deployed for the clinical entity recognition independently. To take the advantages of different methods, we use a vote-based approach to combine all predicted entities by them: a candidate entity is selected only when it has been exactly predicted by at least two methods.

Except the annotated data, the organizers of 2017 CCKS challenge also provide a set of unlabeled records. To explore the contribution of unlabeled data for the clinical entity recognition, we use a self-training approach, as follow: 1) train all individual methods on the official training set; 2) tag unlabeled records by above methods respectively, and combine all results by voting; 3) merge the tagged unlabeled data with the official training data as a new larger training set; 4) finally, retrain all above individual methods on this new training set, and tag on the official test data.

3 Experiments

3.1 Dataset

In the 2017 CCKS CNER challenge, organizers provided 400 medical records annotated with five categories of entity (as in Table 1), 300 out of them are used as a training set and 100 as a test set. Besides, other 2,605 unlabeled records are released as an unlabeled set. The statistics of entity on different categories are listed in Table 1.

Table 1. Statistics of entity on different categories.

Dataset	Body	Disease	Symptom	Test	Treatment	All
Training set	10,719	722	7,831	9,546	1,048	29,866
Test set	3,021	553	2,311	3,143	465	9,493

3.2 Evaluation

All our evaluations are performed on the official test set using the evaluation tool of 2017 CCKS CNER challenge, which outputs micro-average precisions (Prec.), recalls (Rec.) and F1-scores (F1) under two criteria: “strict” - checks whether the boundary and category of an entity is exactly matched with a gold one; while “relaxed” - only considers the boundary of an entity is overlapped with a gold one of same category, “strict” is the primary one.

3.3 Experiment Setup and Results

In this study, we directly divide the sentences into Chinese characters, which can avoid the boundary error of entity caused by the word segmentation tools. The “BIOES” (B-begin, I-inside, E-end, S-single, O-outside) tags are used to represent the entity. For neural network models, we use the stochastic gradient descent (SGD) algorithm to estimate parameters, and the pre-trained Chinese character embedding was learned from training and unlabeled datasets by word2vec tool. The feature representations are randomly initialized from a uniform distribution ranging in [-1, 1].

Table 2 shows the performance of various methods on test set. We can see that all the machine learning-based methods outperformed the rule-based method, BI-LSTM model achieves much better F1-scores than CRF-based method, the voted results (90.17% under "strict" criterion) outperformed all the individual methods. After applying the self-training approach, the F1-scores of all individual methods are improved by about 1.0% under "strict" criterion, and the F1-score of voted result is improved by 0.06%. However, the BI-LSTM-FEA model performs much poorly than other methods, which performs best on our validation set (divided from training set). We think that the bad results of rule-based method on test set may cause the poor performance of BI-LSTM-FEA model.

Table 2. Results of various methods on test set.

	Models	Strict (%)			Relaxed (%)		
		Prec.	Rec.	F1	Prec.	Rec.	F1
Rule-based		85.89	87.78	86.82	92.24	94.27	93.24
Supervised learning	CRF	91.22	88.20	89.69	95.73	92.57	94.13
	BI-LSTM	90.68	89.67	90.17	95.18	94.12	94.65
	BI-LSTM-FEA	90.47	88.68	89.57	94.91	93.03	93.96
	Voted	94.49	87.79	91.02	97.26	90.37	93.69
Self-training	CRF	92.42	89.09	90.72	96.06	92.60	94.30
	BI-LSTM	91.99	90.30	91.14	95.64	93.88	94.75
	BI-LSTM-FEA	91.80	89.52	90.65	95.80	93.43	94.60
	Voted	92.99	89.25	91.08	96.23	92.36	94.26

Table 3 lists the F1-scores of various self-training methods on each category under the "strict" criterion. We can see that all methods perform much better on "Symptom" and "Test" categories. So, how to extract the body, disease and treatment entities more effectively will be the main focus in our future works. Furthermore, using some external medical dictionaries to improve the word segmentation also is a good idea.

Table 3. F1-scores of various self-training methods on each category (strict %).

Models	Body	Disease	Symptom	Test	Treatment
Rule-based	82.32	70.19	93.94	91.48	67.11
CRF	86.89	77.59	96.51	94.11	77.36
BI-LSTM	87.48	78.97	96.00	94.43	81.47
BI-LSTM-FEA	86.77	78.90	96.23	93.84	79.11
Voted	87.42	78.60	96.34	94.36	78.92

4 Conclusion

In this study, we proposed a hybrid system based on rule, CRF and RNN methods for the entity recognition from Chinese clinical text. Experiments on 2017 CCKS corpus

show that our system achieves the F1-scores of 91.08% and 94.26% under “strict” and “relaxed” criteria respectively, ranking first in this challenge. Among all individual methods, BI-LSTM outperforms rule-based and CRF methods. By applying a self-training approach with unlabeled data, the F1-scores of all machine learning-based methods are improved by about 1.0% under “strict” criterion. The future works of us will focus on the more effective extraction of body, disease and treatment entities.

Acknowledgments

This paper is supported in part by grants: National 863 Program of China (2015AA015405), NSFCs (National Natural Science Foundations of China) (61573118, 61402128, 61473101, and 61472428), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20140627163809422, 20151013161937, JSGG20151015161015297 and JCYJ20160531192358466), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052), Program from the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (93K172016K12) and CCF-Tencent Open Research Fund (RAGR20160102).

References

1. Uzuner, Ö., Solti, I., Cadag, E.: Extracting medication information from clinical text. Journal of the American Medical Informatics Association 17, 514-518 (2010)
2. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association 18, 552-556 (2011)
3. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association 20, 806-813 (2013)
4. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J.: Overview of the ShARe/CLEF eHealth evaluation lab 2013. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 212-231. Springer, (Year)
5. Pradhan, S., Elhadad, N., Chapman, W.W., Manandhar, S., Savova, G.: SemEval-2014 Task 7: Analysis of Clinical Text. In: SemEval@ COLING, pp. 54-62. (Year)
6. Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., Verhagen, M.: SemEval-2015 Task 6: Clinical TempEval. In: SemEval@ NAACL-HLT, pp. 806-814. (Year)
7. Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., Verhagen, M.: Semeval-2016 task 12: Clinical tempeval. Proceedings of SemEval 1052-1062 (2016)
8. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., Johnson, S.B.: A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association 1, 161-174 (1994)
9. Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., Xu, H.: Entity recognition from clinical texts via recurrent neural network. BMC Medical Informatics and Decision Making 17, 67 (2017)