A Conditional Random Fields Approach to Clinical Name Entity Recognition

Xiaoran Yang and Wenkang Huang*

Alibaba Health Information Technology Limited
{ xiaoyang.yxr,wenkang.hwk}@alibaba-inc.com

Abstract. Clinic named entity recognition (CNER) is an initial step in understanding and using electronic medical record clinical free-text. The CCKS committee sets up a task for CNER for recognizing five types of entities including body part, independent symptom, symptom description, operation and drug. For this task, we develop a conditional random fields (CRF) model with char embedding, POS, radical, PinYin, dictionary and rule features. Our best model on the test dataset achieves the strict F1-Measure of 0.8926 which ranked the first place.

Keywords: Name Entity Recognition, Electronic Medical Records, NER

1 Introduction

With the growth of the scale of electronic medical records, clinical named entity recognition (CNER) has gradually become an important research topic. The research progress of CNER in China is quite slow due to the lack of uniform standards and public datasets. For this purpose, the CCKS 2018 conference in July 2018, sets up a CNER task to identify entities from Chinese clinical text with a label specification and a training datasets.

Currently, the most effective way to identify named entities is based on machine learning algorithm, such as support vector machines (SVM) [1], conditional random fields (CRF) [2], structured support vector machines (SSVM) [3], recurrent neural network (RNN) with its variant model [4], and convolutional neural network (CNN) with its variant model [5]. In this paper, we participated in the CCKS 2018 CNER task and developed a method based on conditional random fields. By evaluating and choosing a great number of different features in the method including the characteristic feature and feature based on external data, we achieved a F1-Measure of 0.8926 based on the CCKS 2018 CNER task dataset.

* Correspondence: wenkang.hwk@alibaba-inc.com

2 Task Formalism

Clinical named entity recognition task is often considered as a sequence label task. Given a sentence X= <x_1,…, x_n>, the goal is to label each character x_i according the context of X with BMESO (B-Begin M-Middle E-End S-Single O-Outside) notation scheme. The CCKS2018 evaluation task 1 gives the annotation datasets and the unlabeled datasets with 5 pre-defined categories (body part, independent symptom, symptom description, operation and drug). An example of the tag sequence for "患者2个月前因上腹部不适于我院就诊 (the patient went to see a doctor two month ago in our hospital because of epigastrium)" is shown in Figure 1.

患 \O 者 \O 2 \O 个 \O 月 \O 前 \O 因 \O 上 \B-BOD 腹 \M-BOD 部 \E-BOD 不 \B-DES 适 \E-DES 于 \O 我 \O 院 \O 就 \O 诊 \O

Fig. 1. Example of tag sequence.

3 Methods

In this section, we first introduce the CRF method algorithms, then introduce the features used in CRF, including char embedding, POS, radical, PinYin, dictionary and rule.

3.1 Conditional Random Fields(CRF)

A conditional random field (CRF) is a type of discriminative, undirected probabilistic graphical model, which has been widely used for sequence labeling problems. For a given character sequence $z = \{z_1, \dots, z_n\}$ where z_n is the input vector composed of the char and features of ith character, and a given label sequence $y = \{y_1, \dots, y_n\}$ for z. $\gamma(z)$ represent the all of possible labels for z. The CRF model define the formula of the probability of character sequence y with given label sequence z is:

$$p(y|z;\theta) = \frac{\sum_{t=1}^{n} \exp(S(y^{(t)}, z^{(t)}, \theta))}{\sum_{t=1}^{n} \sum_{j \in \gamma(z)} \exp(S(y_j, z^{(t)}, \theta))}$$

Where $S(y^{(t)}, z^{(t)}, \theta)$ are potential function, and θ is the parameters of CRF. In our work, we use the character as a unit for sequence labeling model rather than use the word. Log likelihood function was used to get the loss of the CRF layer. Finally, the viterbi algorithm was used to decode.

3.2 Features

3.2.1 Char Embedding

Given a sequence $X = < x_1, \ldots, x_n >$, distributed embedding vector is used to represent the information for each character. Formally, we look up in a character embedding matrix for embedding vector for each character x_i.

A single English character does not have semantics, while Chinese characters often have strong semantic information. To utilize these semantic information, we use cw2vec [7] instead of word2vec to construct the char embedding matrix. Different from the work of word2vec, it puts forward the concept of "n-gram strokes", which is the semantic structure of the continuous n strokes of Chinese words (or Chinese characters). We have trained a cw2vec model using CCKS2018 training corpus and testing corpus with 128 embedding dims.

3.2.2 Part-of-Speech (POS)

Part-of-speech (POS) features can help identify clinical named entity. For example, body parts are always consist of many nouns, such as “右上腹(the right upper quadrant)” , and a verb often comes before the name of the operation or that of the drug, such as taking a drug or performing a surgery. In this paper, a python library named Jieba was used to implement a POS tagger.

3.2.3 Chinese PinYin

Due to the use of the Pinyin input method, a large number of homophone typos entities have appeared, and these homophone typos entities have not been identified. For example, the "右附件(the right adnexa)" appearing in the text can be identified, but "右附件(the right adnexa)" may be mistakenly written as "有附件(have adnexa)" due to the use of Pinyin input method. These homonym characters cannot be identified. In addition, some similar Chinese characters with the same pronunciation would have the same meanings .Therefore, we use character spell features to help improve the result of clinical named entity recognition.

3.2.4 Radical

Chinese characters are composed of smaller units - radicals, like English words are composed of letters. These radicals often have semantic information about the original character. For example, the characters “肠(intestines)”, “肺(lung)”, “肝(liver)” with same radical “月” are all related to human body parts. We retrieved the radical composition of each character from online Xinhua dictionary (http://tool.httpcn.com/Zi).

3.2.5 Dictionary

An additional dictionary was constructed from the training set and open websites or databases such as DrugBank, “xunyiwenyao”, etc. Bi-directional maximum matching (BDMM) algorithm [8] was used to find the word in dictionary appearing in sequences. In order to improve the accuracy of entity boundary recognition, BMESO notation schema was used for tagging which can give more information about character’s position.

3.2.6 Rule

With these dictionaries above, by mining frequent pattern [9], we can also find many medical terminology do not appear in the dictionary. For instance, according the sequence "行子宫切除术(do hysterectomy)，" and "子宫切除术(hysterectomy)" in operation dictionary, we can extract the pattern "行(do)<Operation>，". Using the pattern we can also extract operation entity "直肠癌切除术(rectal cancer resection)" from "行直肠癌切除术(do rectal cancer resection)"，while "直肠癌切除术(ectal cancer resection)" are not in operation dictionary. We also use body part prefixes to extend the body part entities such as "左侧卵巢(the left ovary)" while only "卵巢(ovary)" in body part dictionary. In this paper, words that extracted by patterns were also tagged using BMESO notation schema.

4 Experiments

4.1 Datasets

The CCKS 2018 CNER task provided 600 annotated corpus as training dataset with five types of entities (body part, independent symptom, symptom description, operation and drug). 400 unlabeled corpus were also provided as testing dataset to evaluate the model. The statistics of different types of entities in training corpus are listed in Table 1. To choose features and best hyper-parameters, we split 600 training corpus into 480 training corpus and 120 validation corpus.

Table 1. Statistics of entity on different categories in training sets.

Entity	Body	Symptom	Description	Operation	Drug	All
Count	5574	2764	1708	1085	849	11980

4.2 Experimental Settings

By adjusting the hyper-parameters of the training model through the validation datasets, the best hyper-parameters in CRF model was obtained and described below. The model are trained by Adam optimization algorithm [11].

(1) L1 penalty: 1;
(2) L2 penalty: 0.01;
(3) Max iterations: 100;
(4) Epsilon: 1e-5.

4.3 Experiment on CRF Model

In this section, we compare different combinations between six type features in CRF model. The comparative results are listed in Table 2.

Table 2. Comparative results of different combinations in CRF model.

Feature and CRF loss function	F1 in validation sets
Char	0.9140
Char + Char embedding	0.9082
Char + Word segmentation	0.9111
Char + Radical	0.9157
Char + Radical + POS	0.9160
Char + Radical + POS + PinYin	0.9180
Char + Radical + POS + PinYin + Dictionary	0.9510
Char + Radical + POS + PinYin + Dictionary + Rule	0.9729

The result of CRF model has improved a little with radical features, POS features, and PinYin features, but with dictionary features and rule features, it has improved notably. It seems that radical features, POS features, PinYin features may have potential influence in clinical named entity recognition, but dictionary features and rule features could have explicit improvement.

4.4 Compared with the state-of-art model

In this section, we compare the best CRF model with a state-of-art model bi-LSTM-CRF by testing sets. The comparative results are summarized in Table 3.

Table 3. Comparative best results between two models in test sets.

Model	Evaluation	Body	Symptom	Description	Operation	Drug	All
Bi-LSTM+CRF	Strict	0.8812	0.9184	0.8994	0.8506	0.9343	0.8897
	Relaxed	0.9572	0.9522	0.9220	0.9310	0.9516	0.9511
Our CRF	Strict	0.8797	0.9245	0.9059	0.8543	0.9449	0.8913
	Relaxed	0.9556	0.9552	0.9304	0.9325	0.9620	0.9522

Compared strict and relaxed results, we find that the body parts and the operations don't have a high strict F-measure but have a high relaxed F-measure. It means that the right position of entities has been found without the right boundary. Through searching the full testing corpus, it seems that the body part and operation entities are lack of a uniform labeling specification.

Comparing results between two models, the reason that why the best result in CRF model is better than it in Bi-LSTM-CRF model may be the scale of the data sets smaller than the scale of the entities. Therefore, the Bi-LSTM-CRF model is easy to fall into overfitting. And by looking through the result, we can find that Bi-LSTM-CRF model can identify more entities while some of them are wrong. We believe that if the scale of data sets become larger, the result of Bi-LSTM-CRF will be better.

5 Conclusion

By building a number of features including characteristic of character and external data, a clinical named entity recognition model using CRF algorithm was developed. Compared with the state-of-art algorithm Bi-LSTM+CRF, our CRF model achieved a better performance. The reason might be that the scale of corpus is not large enough and the label specification is not uniform. In the CCKS 2018 CNER task, we achieved a strict F-measure of 0.8926 which ranked the first. We will focus on the more effective extraction of body and operation entities' boundary in the future.

References

1. Asahara Masayuki, and Yuji Matsumoto.: Japanese named entity extraction with redundant morphological analysis. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 8-15 (2003).
2. McCallum Andrew, and Wei Li.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 188-191 (2003).
3. Lee Yuh-Jye, and Olvi L. Mangasarian.: SSVM: A smooth support vector machine for classification. Computational optimization and Applications 20.1, 5-22 (2001).
4. Huang Zhiheng, Wei Xu, and Kai Yu.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv, 1508.01991 (2015).
5. Strubell Emma, et al.: Fast and accurate entity recognition with iterated dilated convolutions. arXiv preprint arXiv, 1702.02098 (2017).
6. Xu Yan, et al. "Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries." Journal of the American Medical Informatics Association 21.e1, e84-e92 (2013).
7. Cao Shaosheng, et al.: cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information. (2018).
8. Gai Rong Li, et al.: Bidirectional maximal matching word segmentation algorithm with rules. Advanced Materials Research. Vol. 926. Trans Tech Publications, 3368-3372 (2014).
9. Xu Dong, et al.: Data-driven information extraction from Chinese electronic medical records. PloS one 10.8, e0136270 (2015).
10. Gross, Samuel S., et al.: Training conditional random fields for maximum labelwise accuracy. Advances in Neural Information Processing Systems (2007).
11. Kingma Diederik P., and Jimmy Ba.: Adam: A method for stochastic optimization. arXiv preprint arXiv,1412.6980 (2014).