Named Entities Recognition for Chinese Clinical Texts by CRF Models

Junjie Luo

School of Management and Economics, The Chinese University of Hong Kong, Shenzhen
Longxiang Avenue 2001, Longgang, Shenzhen, China P. R.

floydjjluo@outlook.com

Abstract. Clinical named entity recognition (NER) that identities boundaries and types of medical entities is a fundamental and crucial task in clinical natural language processing and healthcare. Machine learning methods have been commonly utilized to carry out this task, such as CRF, LSTM. In this work, I propose a framework which could train both one-layer and two-layer CRF models with selected features. The proposed system has been evaluated in CCKS2017 task2 data and achieved 92.41% F1-score for NER task, which performed better than the best F1-score achieved by task participants.

Keywords: Named Entity Recognition, Chinese Clinical Text, Conditional Random Fields

1 Introduction

Electronic medical record (EMR) systems have been widely adopted in China. In the field of healthcare, many tasks in clinical text mining rely on accurate clinical named entity recognition (NER). Traditionally, most of the effective NER approaches are based on machine learning techniques, such as SVM, HMM, CRF, LSTM. For NER task, existing efforts include rule or dictionary-based methods and supervised methods. In this paper, I treat NER as a sequence labeling problem, and use CRF model to address this problem, and develop a system which achieved 92.41% F1-score.

2 Dataset

The annotated clinical text data is provided by China Conference on Knowledge Graph and Semantic Computing 2017 (CCKS2017) for its clinical texts NER task. The data includes: general items (一般项目), medical history (病史特点), diagnosis & treatment (诊疗经过), discharge summary (出院情况), There are 5 types of clinical entities to be recognized, including body, symptom, disease, check, and treatment. Table 1 shows the brief statistic of this dataset.

Table 1. *Brief Statistic of the Annotated Dataset*

	body	symptom	treatment	disease	check
general items (400)	248	758	2	84	2
medical history (400)	8144	5972	253	938	7814
diagnosis & treatment (399)	1185	642	1249	249	1152
discharge summary (399)	4163	2770	9	4	3721
sum	13740	10142	1513	1275	12689

3 Methodology

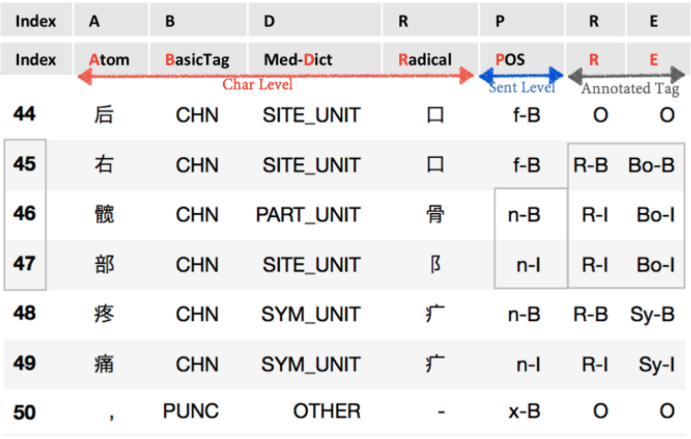
3.1 Character Representation

Due to the limitations of the existing word segmentation tools in Chinese clinical medical text, here I cut the medical text based on the Chinese single character directly. I convert all the annotated texts into the following data structure. Every column represents a character as well as its attributes and annotated information. For example, in text of "外伤后右髋部疼痛, " (After an open wound, the right hip pains,) , "右髋部" (right hip) and "疼痛" (pain) are annotated as body and symptom. This annotated text is processed into Table 2.

3.1.1 Tagging Scheme for Annotation Information

As showed in Table 2, in the processed text data, column *R* and column *E* contain the annotated information. *R* contains entities without type, while *E* contains entities of specific type. I use "R" as an entity tag to represent entity without type, and use "Sy", "Bo", "Ch", "Tr", and "Di" as entity tags to represent symptom, body, check, treatment, and disease. Besides, I utilize following tags of entity boundaries: B, I, and O, which respectively represent a character at the beginning, inside (including the end) or outside of an entity. Entity tags and boundary tags are combined together as final tags. For example, in Table 2, the tags of "右髋部" (right hip) in are "R-B, R-I, R-I", and in are "Bo-B, Bo-I, Bo-I".

Table 2. *Processed Annotated Text Data*



Index	A	B	D	R	P	R	E
Index	Atom	BasicTag	Med-Dict	Radical	POS	R	E
	Char Level				Sent Level	Annotated Tag	
44	后	CHN	SITE_UNIT	口	f-B	O	O
45	右	CHN	SITE_UNIT	口	f-B	R-B	Bo-B
46	髋	CHN	PART_UNIT	骨	n-B	R-I	Bo-I
47	部	CHN	SITE_UNIT	阝	n-I	R-I	Bo-I
48	疼	CHN	SYM_UNIT	疒	n-B	R-B	Sy-B
49	痛	CHN	SYM_UNIT	疒	n-I	R-I	Sy-I
50	，	PUNC	OTHER	-	x-B	O	O

3.1.2 Feature Extraction

The extracted features of each character are divided into two levels. The first level is the character level, including *Atom*, *BasicTag*, *Med-Dict*, *Radical*.

Atoms represent the character in the texts.

BasicTags represent atoms' basic categories, such as CHN (Chinese), ENG (English), PUNC (Punctuation).

Med-Dict represents some medical information contained in each character, such as PART_UNIT (body part), SITE_UNIT (left, right, etc.).

Radical are Chinese radical of Characters. Chinese radical can reflect some information of corresponding characters, such as some body-related characters (脸-face, 脚-foot, 腿-leg) contain radical "月".

The second level is the sentence level, which currently only includes *POS* (Part-of-Speech) tags. I also use "BIO" tagging scheme here to make final *POS* tags. For example, in Table 2, 髋部 (hip) is recognized as n (noun), so I tag "髋部" as "n-B, n-I".

3.2 Models

3.2.1 Conditional Random Field

Conditional Random Fields (CRFs) are undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine.

Let $O = [o_1, o_2 \dots o_n]$ be a sequence of observed words of length n. Let be a set of states in a finite state machine, each corresponding to a label $l \in L$.

Let $S = [s_1, s_2, \dots s_n]$ be the sequence of states in that correspond to the labels assigned to words in the input sequence o.

Linear chain CRFs define the conditional probability of a state sequence given an input sequence to be:

$$P(s|o) = \frac{1}{Z_0} \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(s_{i-1}, s_i, o, i)\right)$$

where Z_0 is a normalization factor of all state sequences. $f_i(s_{i-1}, s_i, o, i)$ is one of m functions that describes a feature, and λ_j is a learned weight for each such a feature function. These weights are set to maximize the conditional log likelihood of labeled sequences in a training set $D = [(o, l)_1, \dots (o, l)_n]$:

$$\log L(D) = \sum_{i=1}^{n} \log\big(P(l_i|o_i)\big) - \sum_{j=1}^{m} \frac{\lambda_j^2}{2\sigma^2}$$

When the training state sequences are fully labeled and unambiguous, objective function is convex, thus the model is guaranteed to find the optimal weight settings in terms of $\log L(D)$. Once these settings are found, the labeling for a new, unlabeled sequence can be done by using a modified Viterbi algorithm. Details of CRFs were completed by Lafferty et al. (2001).

To utilize CRF, I need to adjust model's features and configurations. In this paper, besides the features discussed in *3.1.2*, I set length of the window as 5 for each character, and then extract features with 1-gram, 2-gram, and 3-gram.

3.2.2 One-Layer CRF

The work flow of one-layer CRF model is showed in Figure 1. The target model is *1abdp*, which means the model is a one-layer CRF model with selected features: *a* (*Atom*), *b* (*BasicTag*), *d* (*MedDict*), and *p* (*POS*).

First, the whole dataset keeps the selected features and column *E*, and then is divided into train set and test set in the proportion of 3:1 randomly. For train set, I use *a*, *b*, *d*, *p* to predict *E* by CRF++, and get a trained model instance. For test set, I first derive a golden test set, then use the trained model and test set to obtain the predicted result. By comparing the predicted result with golden test set, I get model *1abdp*'s precision rate, recall rate, and F1 rate.

3.2.3 Two-Layer CRF

Another model is two-layer CRF, whose workflow is showed in Figure 2. The target model is *2abp*, which means the model is a two-layer CRF model with selected features: *a* (*Atom*), *b* (*BasicTag*), and *p* (*POS*).

First, the whole dataset keeps the selected features and both column *R* and column, and then is divided into train set and test set in the proportion of 3:1 randomly. For train set, I use *a*, *b*, *p* to predict column *R*, and get model1, then use *a*, *b*, *p*, *R* to predict column *E*, and get model2. For test set, I first use *a*, *b*, *p* and model1 to get predicted *R*, then use *a*, *b*, *p*, predicted *R* and model2 to get predicted *E*. The same evaluation method used in one-layer CRF are implemented to get precision rate, recall rate, and F1 rate for model *2abp*.

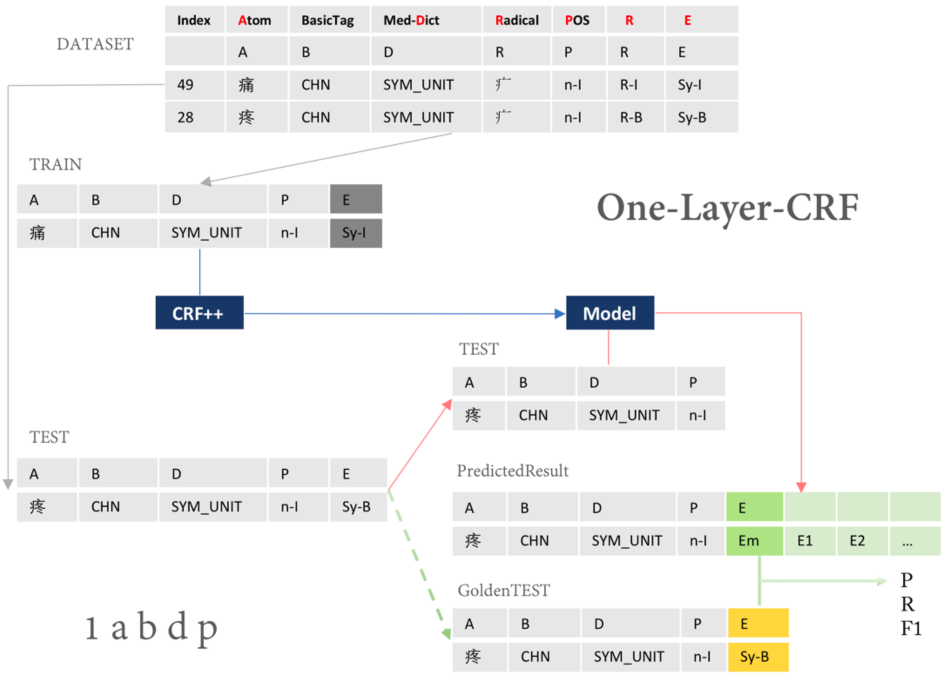


Figure 1. *Workflow of One-Layer CRF Model Structure*

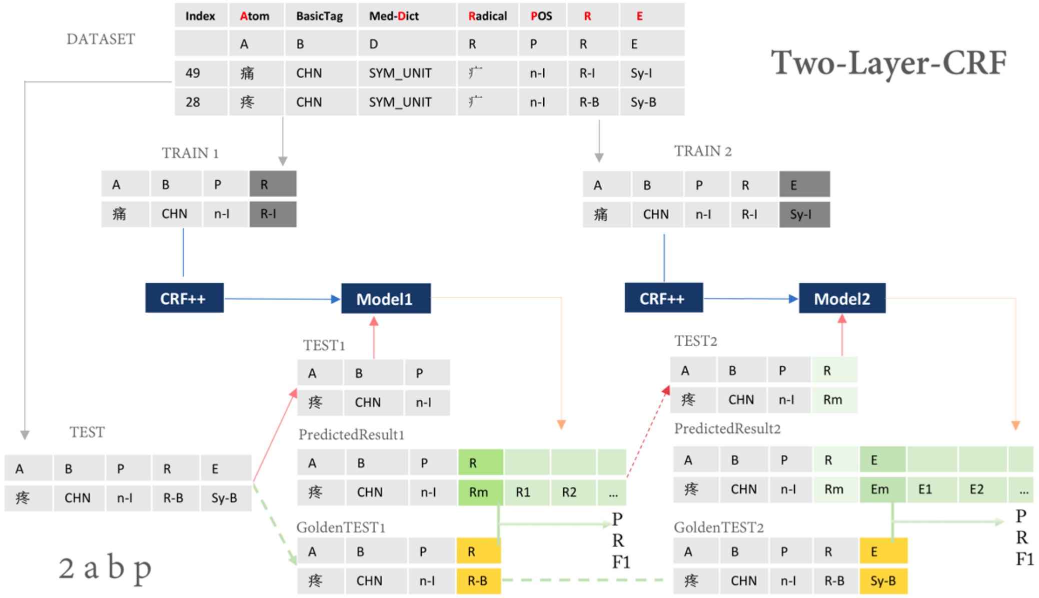


Figure 2. *Workflow of Two-Layer CRF Model Structure*

3.3 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 the same mention, boundaries, and entity type, while the 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 via the strict metrics.

4 Implementation and Experiment Results

4.1 Implementation tools

The whole system is implemented using Python 3.6.2. Chinese word segmentation and POS tagging are implemented using Jieba 0.38. CRF model is implemented by CRF++ 0.58. This system's documents and tutorials are developed by Jupyter Notebook. The source code is available in: https://github.com/floydluo/cctner.

4.2 Experimental Results

In this work, 13 models are trained. The F1 results are listed in Table 3. The last three rows are the models' results from other groups who attended CCKS2017 task2. These three rows show the best performances of each model. For example, the last row shows the best performance of all Bi-LSTM models. LSTIM-CRF result is from Xi, et al (2017), while CRF and Bi-LSTM results are from Hu, et al (2017).

Table 3. *Experiment F1 Results and Best Results achieved by Task Participants*

	body	symptom	treatment	disease	check	overall
1ab	89.11%	97.20%	73.89%	76.31%	95.03%	92.30%
1ar	89.07%	97.09%	74.63%	75.73%	94.84%	92.21%
1ap	**89.40%**	97.20%	75.32%	75.53%	94.74%	92.34%
1abr	88.65%	**97.40%**	**76.86%**	78.17%	95.28%	**92.41%**
1abp	88.71%	97.37%	76.01%	77.12%	95.06%	92.29%
1adr	88.13%	97.24%	76.82%	**79.90%**	95.11%	92.18%
1abdp	88.96%	97.32%	76.69%	77.84%	94.96%	92.38%
2ab	88.77%	96.50%	71.58%	73.60%	**95.36%**	91.93%
2ar	88.55%	96.71%	72.20%	73.20%	95.21%	91.86%
2ap	89.12%	96.74%	72.65%	77.86%	94.88%	92.11%
2abr	88.29%	96.72%	73.38%	76.14%	95.08%	91.85%
2abp	88.49%	96.72%	75.11%	78.37%	94.94%	92.01%
2arp	88.24%	96.96%	75.10%	77.75%	94.75%	91.90%
LSTM-CRF	83.61%	95.07%	75.51%	76.10%	93.19%	88.85%
CRF	86.89%	96.51%	77.36%	77.59%	94.11%	90.72%
Bi-LSTM	87.48%	96.00%	***81.47%***	78.97%	94.43%	91.14%

4.3 Analysis of Results

The models trained from my system perform well in this entity recognition task. From table 3, we can find my models perform much better than models of best performances in terms of overall F1 rate (92.41% vs 91.14%). For the five entities, except treatment, all the entities predicted by models from my system are of higher F1 rate. Considering my train dataset and test dataset are different from those groups, models from my systems perform as better as, if not better than, these models of best performances.

Results vary significantly across different entities, symptom's and check's F1 rates are more than 95%, and body's F1 rate is near 90%. However, the models' performances on treatment and disease are poor, whose F1 rates are less than 80%. An important reason for this is the limited numbers of annotated treatment and disease entities, whose numbers are nearly 1/10 of the other entities' numbers, as listed in Table 1. Therefore, to extract disease, treatment entities more effectively, we need more annotated texts. For this issue, the result may be better if a rigorous medical dictionary is included.

Besides, within a certain entity, the performances are consistent across different models, including models trained in my systems and models of other groups, which indicates these models, complicated as LSTM-CRF, or naïve as *1ab*, works basically at the same level. This convergence may result from the limited trained data. Currently, the best model is *1abr* (one-layer CRF with features including *Atom*, *BasicTag*, *Radical*), whose F1 is 92.41%. Compared with different CRF model structures' performances, surprisingly, the one-layer CRF models outperform the two-layer CRF models.

Furthermore, the high-quality and consistent annotation guideline is crucial in NER tasks. Some errors are founded when I investigated the annotated dataset. First, some entities contain an empty space, such as " 右下腹痛" in general item group, "颈 椎" in diagnosis & treatment group, which interfere CRF++'s work significantly. Second, there are entities without annotated in some EMR files while annotated in other files, such as "神经" (nerve) was annotated as body in some files, while it was not annotated in other files. Third, the annotation strategy of complex entities may be inconsistent. For instance, some files annotated "心肺腹查体" (body check for heart, lung, stomach) as a check entity, while some files separated it as "心" (heart), "肺" (lung), "腹" (stomach) as *body* entities, and "查体" (body check) as a *check* entity. In order to achieve higher performance, a new annotation guideline with a higher quality and consistency are required.

5 Conclusion

This work presents a system which could complete Chinese clinical named entity recognition task. The system is based on machine learning completely, without using neither rule-based method or post-processing module. I developed a system which could train CRFs models with different selected features either by using one-layer CRF or two-layer CRF. In the named entity recognition task, I achieved an F-1 score of 92.41%, which outperformed the best results achieved by the task participants.

Acknowledgements

These medical records used in this paper were provided by Jims-Cloud, and CCKS2017 Tasks were supported by the following Grants: Tsinghua Engineering Group, MSRA. Thanks to the organizing committee of CCKS and the annotators of the dataset.

References

Hu, J.L., Shi, X., Liu, Z.J., Wang, X.L., Chen, Q.C., & Tang, B.Z. (2017). HITSZ_CNER: A hybrid system for entity recognition from Chinese clinical text. *China Conference on Knowledge Graph and Semantic Computing 2017.*

Lafferty, J.D., McCallum, A., & Pereira C.N.F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the 18th International Conference on Machine Learning*, (pp. 282-289).

Ouyang, E., Xi, Y.X., Jin, L., Li, Z.F., & Zhang, X.Y. (2017). Exploring N-gram Character Presentation in Bidirectional RNN-CRF for Chinese Clinical Named Entity Recognition. *China Conference on Knowledge Graph and Semantic Computing 2017.*