Named Entity 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 yet crucial task in clinical natural language processing and healthcare. Machine learning methods, such as CRF, LSTM, have been commonly utilized to conduct this task. This paper proposes a framework which can train both one-layer and two-layer CRF models with selected linguistic information. These models can extract five entity types from Chinese clinical texts, such as *symptoms* and *disease*. The proposed framework is evaluated with CCKS2017 task2 data and achieves 92.41% F1-score, which outperforms the best F1-scores achieved by other 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 EMRs text mining rely on accurate clinical named entity recognition (NER). Traditionally, most of effective NER approaches are based on machine learning algorithms, such as support vector machine (SVM), hidden Markov model (HMM), conditional random field (CRF), and long short term memory (LSTM). For the 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 models to address this problem. A framework, which can train both one-layer and two-layer CRF models with different linguistic information combinations, is developed. This framework is evaluated with annotated Chinese clinical texts provided by CCKS2017 for the task of recognizing *body*, *symptom*, *treatment*, *disease*, and *check* from free Chinese clinical texts. The models learned from this framework achieved 92.41% overall F1-score, and most of them outperformed the best result achieved by other task participants.

2 Dataset

The annotated clinical text data is provided by China Conference on Knowledge Graph and Semantic Computing 2017 (CCKS2017) for its task2: clinical texts NER task. The data includes four groups: general items (一般项目), medical history (病史特点), diagnosis & treatment (诊疗经过) and discharge summary (出院情况). There are 5 types of clinical entities to be recognized, including *body*, *symptom*, *treatment*, *disease* and *check*. 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

From the Table 1, we find that: (1) the number of annotated entities are not evenly distributed across the groups. Take entity *treatment* as an example, there are only 2 annotated *treatment* entities in general items, while 1249 in diagnosis & treatment. (2) the total numbers of each annotated entity are different from each other significantly. For example, there are totally 13,740 annotated *body* entities, while only 1,275 *disease* entities. The total number of each entity will influence the models' recognition performances for that entity, which will be discussed later in this paper.

3 Methodology

3.1 Representation of Character Information

This paper treats this NER task as a sequence labeling problem. To solve this, we need to represent the annotated clinical texts as the observations and corresponding states in the sequence. Character-based and word-based representations are two common methods. 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. In order to conveniently select the interested information to train models, I process all the annotated texts into a table format. This table contains the linguistic and annotated information of each character. For example, for the text "外伤后右髋部疼痛, " (after an external wound, the right hip pains), where "右髋部" (right hip) and "疼痛" (pain) are annotated as *body* and *symptom*, it is processed as Table 2. Next, I will talk about the tagging schemes for representing linguistic and annotated information.

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

3.1.1 Tagging Scheme for Linguistic Information

Each character itself is an observation in the sequence. Intuitively, adding its additional information will improve models' recognition performances. To get each observation's additional information, this paper extracts each character's linguistic information and represents it with a tagging scheme.

The linguistic information is divided into two levels: (1) character level. It includes *Atom*, *BasicTag*, *Med-Dict*, *Radical*. *Atom* represents the character itself in the texts. *BasicTag* represents the character's basic category, such as CHN (Chinese), ENG (English) or PUNC (punctuation). *Med-Dict* represents some medical information contained in the character, such as PART_UNIT (body part), SITE_UNIT (left, right, etc.). *Radical* is the character's Chinese radical. Chinese radicals can reflect some information of corresponding characters, such as some body-related characters (脸-face, 脚-foot, 腿-leg) contain radical "月". (2) sentence level. It currently only includes part-of-speech (POS) tags. To generate POS tags for each character, "BIO" tagging scheme is also used here as entity boundary tags, where B, I, and O represent a character at the beginning, inside (including the end) or outside of an entity, respectively. For example, in Table 2, 髋部 (hip) is recognized as a n (noun), so I tag "髋部" as "n-B, n-I".

3.1.2 Tagging Scheme for Annotated Information

Besides the linguistic information, the processed annotated texts data also contains the annotated information. As showed in Table 2, column *R* and column *E* represent the annotated information. *R* represents entities without types, while *E* represents entities of specific types. 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 combine these entities tags with "BIO" entity boundary tags to get final tags. For example, in Table 2, the tags of annotated *body* entity "右髖部" (right hip) are "R-B, R-I, R-I" in column *R* and "Bo-B, Bo-I, Bo-I" in column *E*.

In this paper, the relationship between linguistic information and annotated information will be learned to train models. Based on the linguistic information of characters from a new free Chinese clinical text, these models are expected to predict the corresponding entity labels for each character.

3.2 Model and Implementation

3.2.1 Conditional Random Field Model

To learn relationship between linguistic information and annotated information, this paper uses Conditional Random Fields (CRFs). CRFs are undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine. This model is suitable to solve sequence labeling problems. CRFs have been shown to be efficient in many natural language processing tasks, such as POS tagging and NER.

Let $o = [o_1, o_2, \dots o_i, \dots o_n]$ be a sequence of observed characters of length n. For example, o_i is a character, such as "痛" (hurt). Let S be a set of finite states. Let $s = [s_1, s_2, \dots s_i, \dots s_n]$ be the sequence of states in S that correspond to the tags assigned to characters in the input sequence o. For example, s_i corresponds to an entity tag, such as "Sy-I". This paper utilizes linear chain CRFs that use a first order Markov independent assumption with binary feature functions. Linear chain CRFs define the conditional probability of a state sequence s given an observation sequence o 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, $f_j(s_{i-1}, s_i, o, i)$, which describes a feature, is one of m feature functions, and λ_j is a learned weight for each feature function. For the train set $D = [(o_1, s_1), (o_2, s_2), \dots (o_n, s_n)]$, these weights are learned to maximize the conditional log likelihood of labeled sequences:

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

When state sequences of the train set 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. More complete details of CRFs were presented by Lafferty et al. (2001).

In this paper, the software CRF++ is used to train the model based on processed annotated texts. Configurations of features are needed to fully utilize CRF models. This paper sets length of the window as 5 for each character, and then extract features with 1-gram, 2-gram, and 3-gram.

3.2.2 Implementation of One-Layer CRF

This paper designs two model structures: one-layer CRF and two-layer CRF. These structures implement CRF models for learning the relationship between linguistic information and annotated information. This section uses a model instance *1abdp* to illustrate how the one-layer CRF works. The name *1abdp* indicates a one-layer CRF model with selected linguistic information: *a* (*Atom*), *b*

(*BasicTag*), *d* (*Med-Dict*) and *p* (*POS*). The workflow of training and evaluating model *1abdp* is showed in Figure 1. First, the whole dataset keeps the selected linguistic information (*a*, *b*, *d*, *p*) and annotated information (column *E* only). Then, it is divided into train set and test set randomly in the proportion of 3:1. For train set, CRF++ learns relationship between *a*, *b*, *d*, *p* and *E* to generate a model instance. For test set, linguistic information (*a*, *b*, *d*, *p*) is fed into the learned model instance to predict the entity tags. Then, by comparing the predicted entity tags with annotated information in test set, I get model *1abdp*'s precision rate, recall rate and F1-score.

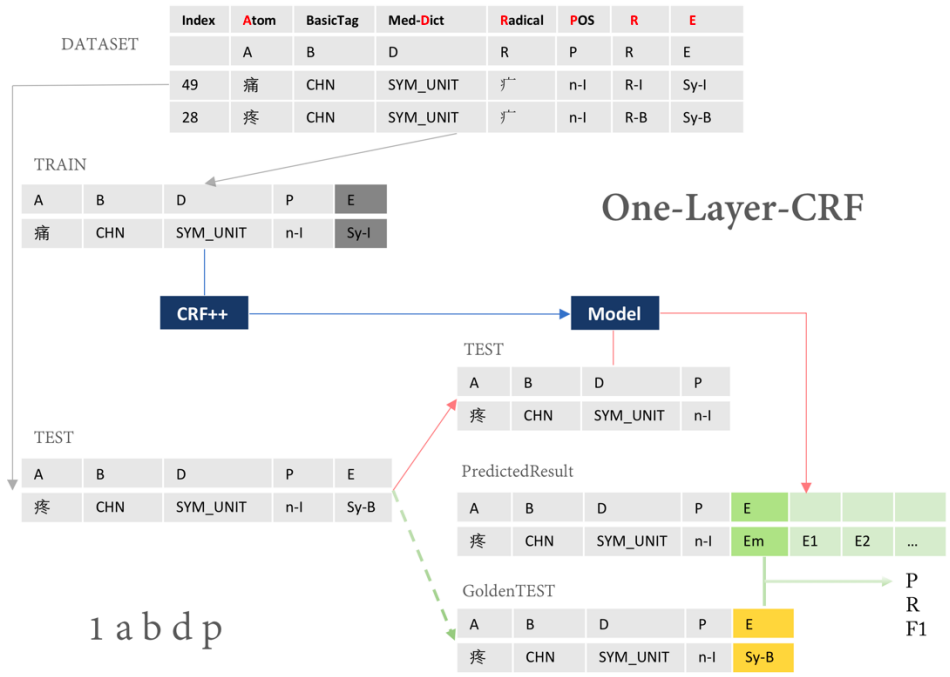

Figure 1. *Workflow of One-Layer CRF Model Structure*

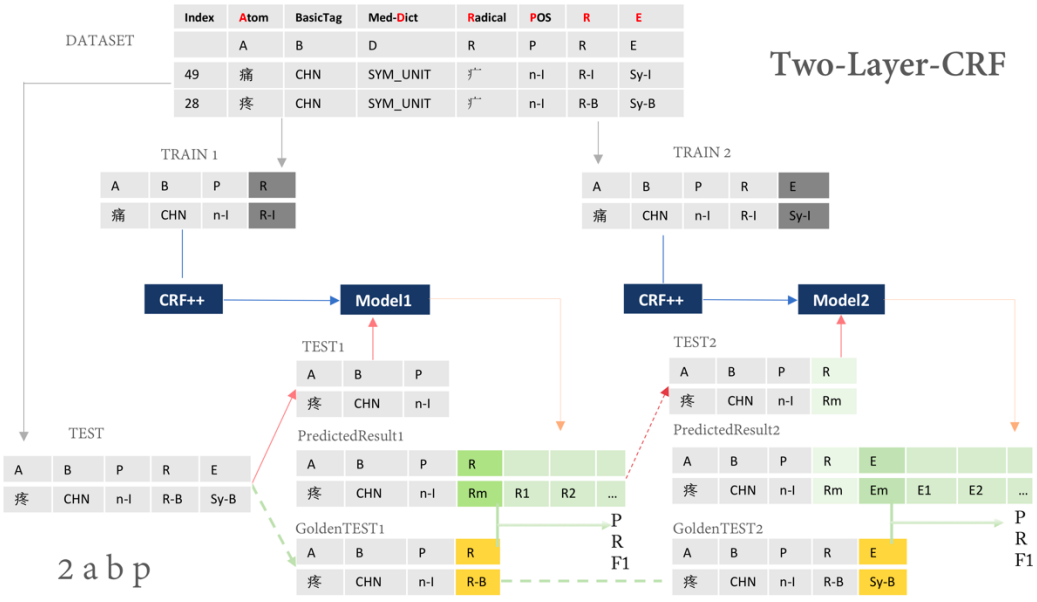

Figure 2. *Workflow of Two-Layer CRF Model Structure*

3.2.3 Implementation of Two-Layer CRF

Another structure is the two-layer CRF, whose workflow is showed in Figure 2. The target model's name is *2abp*, which means the model is a two-layer CRF model with selected linguistic information: *a* (*Atom*), *b* (*BasicTag*) and *p* (*POS*). First, the whole dataset keeps the selected linguistic information (*a*, *b*, *p*) and all annotated information (both column *R* and column *E*). Then it is divided into train set and test set randomly in the proportion of 3:1. For train set, CRF++ learns relationship between *a*, *b*, *p* and column *R* to generate model1, and then learns relationship between *a*, *b*, *p*, *R* and column *E* to generate model2. For test set, linguistic information (*a*, *b*, *p*) is fed into model1 to predict *R* entity tags. Then, linguistic information (*a*, *b*, *p*) and predicted *R* tags are fed into model2 to predict *E* tags. Finally, the same evaluation method is used to get model *2abp*'s precision rate, recall rate, and F1-score.

3.3 Evaluation Measures

The evaluation metrics of this task include: (1) strict metric. It defines a correct match as that the ground truth and extraction result share the same mention, boundaries and entity type. (2) relaxed metric. It defines a correct match as that the ground truth and extraction result share same entity type and overlap boundaries. This paper evaluates models' F1-score performances via the strict metric.

4 Implementation and Experiment Results

4.1 Implementation tools

The whole system is developed by 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 system's source code is available in: https://github.com/floydluo/cctner.

4.2 Experimental Results

In this work, 13 model instances are learned. Their F-score results are listed in Table 3. The last three rows are F1-score results from the models of other participants who attended CCKS2017 task2. These three rows show the best performances of each model type. 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. *F1-score Results and Best Results Achieved by Other 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%

Besides, a system that can extract these five clinical entities (*body*, *symptom*, *treatment*, *disease* and *check*) from free Chinese clinical texts is developed based on these learned models. For the free Chinese clinical texts outside of CCKS2017's annotated texts, this system can also recognize these entities with high correct rates.

4.3 Analysis of Results

The models trained from my system perform well in this entity recognition task. From table 3, we can find that 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 entity types predicted by models from my system are of higher F1-scores. Considered my models' train set and test set are different from that of other participants, models learned by my systems perform as better as, if not better than, the best models developed from other task participants.

Another point is that F1-score results vary significantly across different entities. For example, *symptom*'s and *check*'s F1-scores are more than 95%, and *body*'s F1-score is near 90%. However, the models' performances on *treatment* and *disease* are poor, whose F1-scores are less than 80%. An important reason for this is the limited numbers of annotated *treatment* and *disease* entities, whose numbers are only about 1/10 of the numbers of other entities, as showed in Table 1. Therefore, to extract *disease* and *treatment* entities more effectively, we need more annotated texts. For this issue, the results 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 participants. This indicates that these models, complicated as LSTM-CRF, or naïve as *1ab*, perform basically in the same level. This convergence may be resulted from the limited train data. Currently, the best model is *1abr* (one-layer CRF with linguistic information of *Atom*, *BasicTag*, *Radical*), whose F1 is 92.41%. Compared with different CRF model structures' performances, surprisingly, one-layer CRF models outperform two-layer CRF models.

Furthermore, the high-quality and consistent annotated clinical texts is crucial in NER tasks. Some errors are founded when I investigate the annotated dataset. First, some entities contain an empty space, such as " 右下腹痛" in general item group and "颈 椎" in diagnosis & treatment group, which interfere CRF++'s work significantly. Second, there are entities being annotated in some EMR texts not being annotated in other texts. For example, "神经" (nerve) is annotated as a *body* entity in some texts, but not annotated in other texts. Third, the annotation strategy of complex entities may be inconsistent. For instance, in some texts, "心肺腹查体" (body check for heart, lung and stomach) is annotated as a *check* entity. However, in some other texts, it is treated as annotating "心" (heart), "肺" (lung), "腹" (stomach) as *body* entities, and "查体" (body check) as a *check* entity. In order to achieve high performances, the above errors should be avoided and annotated clinical texts of consistency and a high quality are necessary.

5 Conclusion

This paper presents a system which can 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. This paper develops a system which can train CRFs models with different combination of selected linguistic information by using either one-layer CRF or two-layer CRF. In the named entity recognition task, these models achieve an F-1 score of 92.41%, which outperforms 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.*