PROCEEDINGS OF SCIENCE

Feature Split-based Information Extraction in the Field of Medicine

Jing Wan[1]
Beijing University of Chemical Technology
Beijing, 100029, China
E-mail: wanj@mail.buct.edu.cn

Huanchun Yan
Beijing University of Chemical Technology
Beijing, 100029, China
E-mail: Wendy8Fairy@163.com

Xuechao Zhang[2]
Logistics Academy
Beijing, 100036, China
E-mail: zhangxuechaobj@163.com

In recent years, more and more studies have been done on symptom information extraction. These studies are mostly based on clinical medical records, and they focus only on symptom entities, which are not sufficient to convey the full symptom information. This paper presents a feature split-based approach to extract symptom information from Chinese medicine instruction texts. In this approach, the symptom information is split into two parts: symptom subject entity and symptom manifestation entity. The main idea of this method is to automatically recognize the symptom subject and symptom manifestation first, and then add these two identification results as features to the symptom information extraction task. Through a series of experiments based on Conditional Random Fields (CRF)-an effective model proved by lots of experiments in the field of medicine, it is obvious that the feature split-based approach proposed in this paper can obtain higher accuracy and recall rate in symptom information extraction.

ISCC2017
16-17 December 2017
Guangzhou, China

[1] Speaker

[2] Correspongding Author

© *Copyright owned by the author(s) under the terms of the Creative Commons*
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1.Introduction

Recognition of medical entities can be realized via rule-based methods and statistical machine learning methods. The early named entity recognition (NER) mostly uses rule-based methods, but they are highly dependent on language, field and text style of corpuses with limitation on cost effectiveness. Nowadays, NER researches mostly adopt statistical machine learning methods, such as the common use of HMM [1][2], SVM [3][4], conditional random fields (CRF) [6][7] etc., and more researches indicate that CRF is of favorable effect in NER work [1][3].

International researches on medical entity recognition have set an early start. Zhang S et al verified unsupervised biomedical entity recognition method respectively based on i2b2 and GENIA, and the results indicated that the recognition results of CRF were superior to those of HMM [1]. Liao Z et al, based on long-range dependencies phenomenon between entities in biomedical texts, proposed using skip-chain CRF to recognize medical entities [5]. In 2014 i2b2 de-identification challenge, Yang H et al combined CRF, rules and key words to design a hybrid automatic recognition system, which conquered the challenge by 93.6% of F-measure [10].

In comparison, domestic researches on medical information extraction in China have a late start [9][11] , but have already obtained impressive research achievements. Wang Y et al, based on CRF, studied influences brought by different features on symptom name recognition in traditional Chinese medical records, and their recognition rate and F-measure are respectively 93.403% and 62.829% [9]. Wang Y et al, made a comparative analysis of recognition effects of three machine learning models (CRF, HMM and MEMM) on symptom entities, and then verified that CRF model was more appropriate for symptom entity extraction in traditional Chinese clinical records [2]. Liu H et al studied the performance of different feature templates on symptoms and pathogenesis entity extraction in traditional Chinese clinical records and observed the influence of an increase of training corpuses on CRF performance after the optimal template was determined [8].

Despite of that, the above researches have neglected the influence of selection of annotated set on the acquisition of symptom information and symptom extraction is limited to symptom entities so that they can’t meet the requirements for the acquisition of symptom information during establishment process of medical knowledge base. Symptom entity is not enough to depict full symptom information; hence, it is necessary to understand where the symptoms occur and the occurrence state. As a result, a symptom-information extraction method based on feature splitting and CRF was proposed in this paper. Symptom information was divided into symptom subject and symptom manifestation for separate recognition, and the recognition results were added into symptom information extraction work as features. This method could not only improve precision rate and recall rate of symptom information extraction, but also obtain the parts where symptoms occurred and the occurrence state while recognizing symptom information.

2. Conditional Random Fields

Conditional Random Fields (CRF), as one of the random fields, is a discriminative probabilistic model. It can adopt complicated, overlapping and non-independent features for training and reasoning, also take full advantages of contextual information as features as well as add other external features [12]. In CRF, the distribution of random variable Y is conditional

PoS(ISCC 2017)029

probability, and the given observational value is random variable X. Observational sequence is set as X={X1, X2, X3, …, Xn, where input data can be words, characters or symbols and so on in the text, and its corresponding state sequence is Y={Y1, Y2, Y3, …, Yn}. When the given observational sequence is X, the calculation methods of conditional probability of state sequence and the normalization factor are respectively shown in formula (2.1) and (2.2).

$$p(y|x)=\frac{1}{Z(x)}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\} \qquad (2.1)$$

$$Z(x)=\sum_{y}\prod_{t=1}^{T}\exp\{\sum_{k=1}^{K}\theta_k f_k(y_t,y_{t-1},x_t)\} \qquad (2.2)$$

Where θ_k is weight parameter of corresponding characteristic function; y_t and y_{t-1} are respectively express the present output state and the last output state; x_t is present input state.

3. CRF-based Symptom Information Extraction

3.1 The Framework of Symptom Information Extraction

Symptom information extraction can be regarded as the process of structuring symptom information into dual-tuple structure (P. S), where P represents symptom subject and S represents symptom manifestation. In the first place, symptom subject, symptom manifestation and symptom information of the original corpus are annotated. Secondly, the recognition models of symptom subject and symptom manifestation are respectively obtained based on the learning process of CRF model. To follow that, the annotated results of symptom subject and symptom manifestation are directly added into basic features of symptom information as features. Finally, the symptom information extraction model is obtained on basis of learning process of CRF model. The process of the training model is as shown in Figure 1.

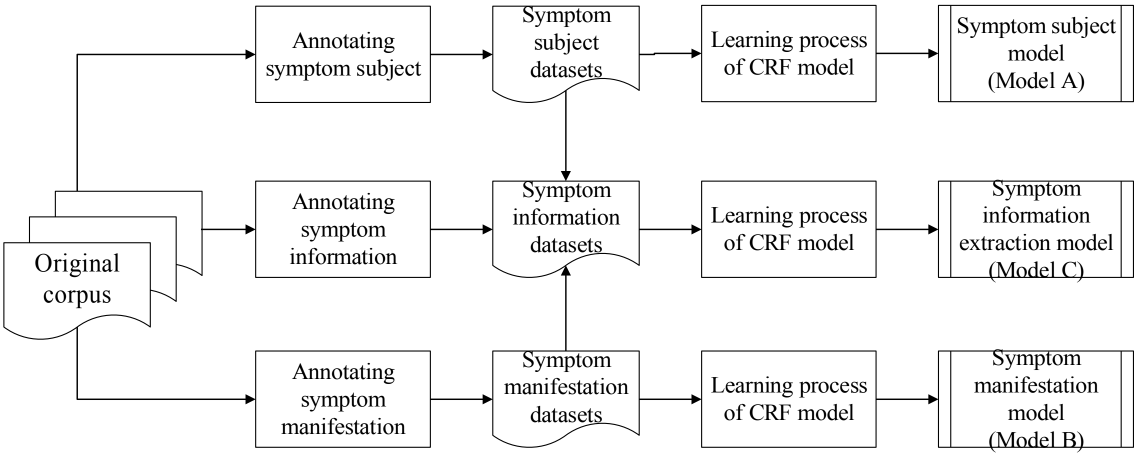

Original corpus
Annotating symptom subject
Symptom subject datasets
Learning process of CRF model
Symptom subject model (Model A)
Annotating symptom information
Symptom information datasets
Learning process of CRF model
Symptom information extraction model (Model C)
Annotating symptom manifestation
Symptom manifestation datasets
Learning process of CRF model
Symptom manifestation model (Model B)

Figure 1: Training Flowchart

While dealing with a new text, symptom subject and symptom manifestation are respectively recognized through the recognition model of symptom subject and symptom manifestation associated with CRF algorithm, then recognition results are directly added into the target text as features. Afterwards, the symptom information is recognized through symptom information extraction model associated with CRF algorithm, and the process is as shown in Figure 2.

PoS(ISCC 2017)029

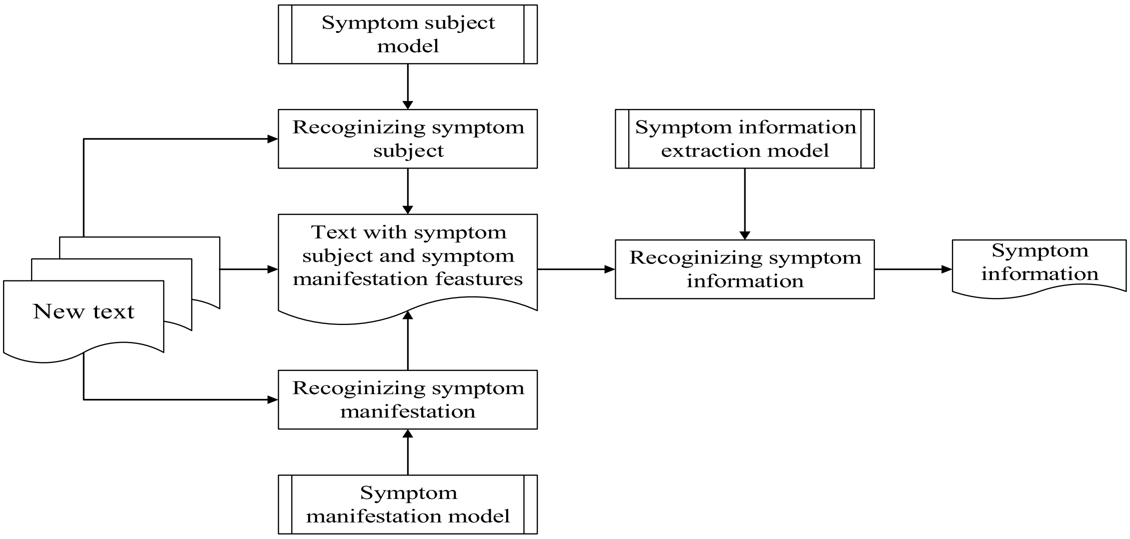

Symptom subject model
Recoginizing symptom subject
Symptom information extraction model
Text with symptom subject and symptom manifestation feastures
Recoginizing symptom information
Symptom information
New text
Recoginizing symptom manifestation
Symptom manifestation model

Figure 2: Symptom Information Extraction Flowchart

3.2 The Features of Symptom Subject and Symptom Manifestation

It's proposed in this paper to split symptom text into symptom subject and symptom manifestation for recognition respectively and add the two recognition results as features of symptom information into extraction model. Symptom subject is namely a subject bearing symptom manifestation and it can be regarded as one body part like "tooth". On the other hand, symptom manifestation is a description of the subject or subject state. The splitting form can be demonstrated by the following form:

SYMPTOMS = SYMPTOMS_PART + SYMPTOMS_STATE.

The splitting process may generate the following circumstances: both SYMPTOMS_PART and SYMPTOMS_STATE exist, for example "headache ", SYMPTOMS_PART is "head " and SYMPTOMS_STATE is "ache"; only SYMPTOMS_STATE exists like "coughing". It is of no significance to independently describe a body part in symptom information, which explains why SYMPTOMS_PART can't independently exist.

3.3 Annotation Strategy

Annotation sets and the explanation used in this research are as shown in Table 1. The annotation set (PART-B, PART-I, POSTPOSITION, O) should be used to train symptom subject recognition model and the other annotation set (STATE-B, STATE-I, POSTPOSITION, O) should be used to train recognition model of symptom manifestation. Annotation set of symptom extraction model is (Symptom-B, Symptom-I, POSTPOSITION, O).

Annotation description	Annotation symbols
Symptom description starts	Symptom-B
In the middle of symptom description	Symptom-I
Subject description starts	PART-B
In the middle of subject description	PART-I
State description starts	STATE-B
In the middle of state description	STATE-I
Non-target data	O

Table 1: Annotation Sets

4.Experiment and Results Analysis

PoS(ISCC 2017)029

4.1 Experimental Dataset

On medical knowledge base, clear and complete record of the parts where symptoms occur and the symptom manifestations contributes to the precision rate of automatic medicine recommendation. Medicine instructions constitute the most fundamental and authoritative textual description of medicine information. So 9,329 medicine instructions texts were taken from YAOZH.com, and 1,500 texts of them were randomly selected as the original data of symptom information extraction experiments in this study.

4.2 Experimental Scheme

The experiments aim to verify the effectiveness of symptom information extraction method proposed in this paper. CRF++0.58 is used as training and test tools. Tenfold cross validation is used in this paper to evaluate the result of symptom information extraction. The evaluation sets base on precision rate, recall rate and F1 measure.

For the convenience of demonstration, features mentioned above are respectively annotated by symbols. Table 2 displays the corresponding relationships between features and English symbols, where PART feature includes PB and PI annotation features and STATE feature includes SB and SI annotation features.

Feature descriptions	Corresponding symbols
Part-of-speech features	POS
Take word as basic unit	WORD
Take character as basic unit	CHAR
Features of symptom subject description start	PB
Features of symptom manifestation description start	SB
Features in symptom subject description	PI
Features in symptom manifestation description	SI
Features of symptom subject	PART
Features of symptom manifestation	STATE

Table 2: Feature Expression Method

The following three groups of experiments are designed to verify the advantages of the method proposed in this paper.

Experiment I: Comparative experiment of character and word-based annotation. The experiment studies extraction effect of annotation method (word or character) on symptom information extraction in original text and influence of adding features of part-of-speech on extraction effect.

Experiment II: Recognition experiment of symptom subject and symptom manifestation. The model trained in this experiment acts as the tool for extracting fetures of symptom subject and symptom manifestation in experiment III.

Experiment III: Feature fusion experiment. The recognition model in experiment II is used to respectively annotate symptom subject and symptom manifestation, the results are added into symptom information extraction model as features, and the influences of the above two features on symptom text recognition are compared.

PoS(ISCC 2017)029

4.3 Experimental Results and Analysis

In the following results, BP, BR and BF respectively represent recognition precision rates, recall rates and F-measure of Symptom-B. Similarly, IP, IR and IF respectively represent recognition precision rates, recall rates and F-measure of Symptom-I.

4.3.1 Comparison of Character and Word-based Annotation

Experimental results are as shown in Figure 3. The experiment indicates that when character is taken as a basic unit of symptom recognition, both precision rate and recall rate are higher than when word is taken as a basic unit. It's not difficult to find through careful observation of original text segmentation results that Chinese segmentation technology is of low precision on segmentation in the field of medicine, with frequent occurrence of wrong segmentation, which will cause fuzzy boundary of symptom information labeling. Consequently, correct results can't be obtained. Hence, the following experiments have given up word-based features of parts of speech, and taken character as a basic unit.

4.3.2 Extraction of Symptom Subjects and Symptom Manifestations

Experimental results of extraction of symptom subjects and symptom manifestations are as shown in Table 3. In this experiment, symptom text is split into symptom subject and symptom manifestation for separate recognition. It can be seen that the extraction results of these two kinds of information are the same in low precision rate and high recall rate. It's noteworthy that this experiment concentrates on the overall extraction effect of symptom information, while as features of recognition, symptom subject and symptom manifestation should be acquired as much as possible. Hence, in symptom subject model and symptom manifestation model, this method pays more attention to recall rate.

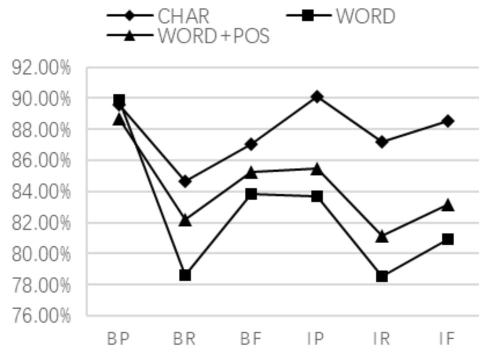
Recognition category	Type	Precision rate (%)	Recall rate (%)	F1 (%)
PART	Symptom-B	31.04	91.65	46.38
	Symptom-I	36.20	73.29	48.46
STATE	Symptom-B	32.59	81.70	46.59
	Symptom-I	37.08	84.46	51.53

Table 3: Recognition Result of Symptom Subject and Symptom Manifestation

4.3.3 Feature Fusion

The experimental results of feature fusion are as shown in Figture 4, after symptom subjects and symptom manifestations are respectively added or simultaneously added, model recognition effect is much better than that when they are not added, which certifies that symptom subject features and symptom manifestation features will benefit symptom information recognition.

PoS(ISCC 2017)029



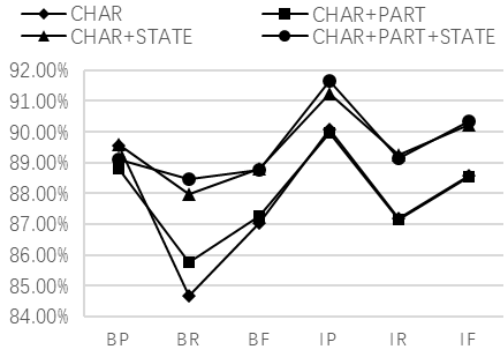


Figure 3: Comparison of Character and Word-based Annotation

Figure 4: Experimental Results of Multi-feature Fusion

PoS(ISCC 2017)029

5.Conclusion

Experimental comparison indicates that the method of symptom information extraction based on text splitting and CRF proposed in this paper can acquire decomposed symptom information and can accomplish higher precision rate and recall rate. Moreover, in the task of symptom information extraction, BIOPost annotation set can significantly enhance extraction precision rate and recall rate on the condition that manual annotation is not added. However, in extraction experiment of symptom subjects and symptom manifestations, the recognition accuracies are comparatively low. In the future, there is the plan to extract localized symptom subjects and symptom manifestation dictionaries to improve recognition precision rate. Furthermore, more features will be studied to optimize the method of extracting symptom information.

References

[1] S. Zhang, N. Elhadad. *Unsupervised biomedical named entity recognition: experiments with clinical and biological texts* [J]. Journal of Biomedical Informatics, 2013, 46(6): 1088-1098

[2] Y. Wang, Z. Yu, L. Chen, Y. Chen, Y. Liu, X. Hu, Y. Jiang. *Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study* [J]. Journal of Biomedical Informatics, 2014, 47(2):91-104

[3] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, H. Xu. *A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries* [J]. Journal of the American Medical Informatics Association Jamia, 2011, 18(5): 601-606

[4] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu. *A comprehensive study of named entity recognition in Chinese clinical text* [J]. J Am Med Inform Assoc, 2013, 21(5): 808-814

[5] Z. Liao, H. Wu. *Biomedical Named Entity Recognition Based on Skip-Chain CRFS* [C]. International Conference on Industrial Control and Electronics Engineering, China, 2012: 1495-1498

[6] L. Yang, Y. Zhou. *Exploring feature sets for two-phase biomedical named entity recognition using semi-CRFs* [J]. Knowledge and Information Systems, 2014, 40(2): 439-453

[7] D. Jain. *Supervised Named Entity Recognition for Clinical Data*. CLEF 2015 Online Working Notes, 2015

[8] H. Liu, X. Qin, B. Fu. *The Symptoms and Pathogenesis Entity Recognition of TCM Medical Records Based on CRF* [C]. IEEE, Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE, Intl Conf on Autonomic and Trusted Computing and 2015 IEEE, Intl Conf on Scalable Computing and Communications and ITS Associated Workshops, China, 2015: 1479-1484

[9] Y. Wang, Y. Liu, Z. Yu, L. Chen, Y. Jiang. *A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features* [C]. The Workshop on Biomedical Natural Language Processing, Canada, 2012: 223-230

[10] H. Yang, J. M. Garibaldi. *Automatic detection of protected health information from clinic narratives* [J]. Journal of Biomedical Informatics, 2014, 79:S30-S38

[11] Y. Wang, Z. Yu, Y. Jiang, Y. Liu, L. Chen, Y. Liu. *A framework and its empirical study of automatic diagnosis of traditional Chinese medicine utilizing raw free-text clinical records* [J]. Journal of Biomedical Informatics,2012, 45(2):210-223

[12] J.Lafferty, A. McCallum, F. Pereira. *Conditional Random Fields: Probabilistic Models For Segmenting And Labeling Sequence Data* [C]. The Eighteenth International Conference on Machine Learning, USA, 2001: 282-289

PoS(ISCC 2017)029