A Method of Naming and Identifying Chinese Medical Cases Based on Multi-feature Template Modification

Shouqiang Chen[1, a],Yang Chen[2, 3, b], Feng Yuan[4, c, *] , Lili Zhao[4, d], Wenrong An[1, e]

[1]*Center of Hear of the Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China*
[2]*School of Information Science and Engineering, Shandong Normal University, Jinan, China*
[3]*University of Waterloo, Ontario, Canada*
[4]*Key Laboratory of TCM Data Cloud Service in Universities of Shandong (Shandong Management University) Jinan, China*
[a]*csq23800@163.com,* [b]*2583952180@qq.com,* [c]*yuanfeng623@163.com,*
[d]*zhll001986@163.com,* [e]*anwenrong@126.com*
**corresponding author*

Keywords: named entity identification, Chinese medical case, conditional random field, multi-feature

Abstract: Chinese medical case is a record of diagnosis and treatment activities of TCM experts. Identification of its named entities is of great significance to standardization and information nalization of Chinese medical cases. In view of vague expression and unclear title in the text of Chinese medical case, based on conditional random fields (CRFs), this paper proposes a named entity identification pat-tern based on multi-feature template modification. Firstly, sentence extraction and automatic word segmentation were performed on the texts of Chinese medical cases. Then, character features, part-of-speech features, left-right designator features and term features were labelled for the corpora after word segmentation. Finally, CRFs models were trained using the labelled data to identify four diagnostic of TCM, syndrome patterns and therapy entities, build triple correspondence of four diagnostic of TCM- syndrome pattern -therapy, and provide reference and basis for scientific argumentation of syndrome differentiation and treatment. With 12,000 Chinese medical cases of cardiovascular outpatient specialists at the Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine as the data source, identification accuracy was further enhanced by different combinations of features and adjustment of context window size. Average accuracy, recall and F measures reached 90.68%, 90.45%, and 90.56%, respectively.

1. Introduction

With the popularization and development of TCM informatization technology, massive Chinese medical cases have grown exponentially. Summarizing the experiences, inheriting the science, conducting entity identification of massive Chinese medical case texts, and establishing the connection be-tween four diagnostic methods of TCM, syndrome pattern and therapy, scientifically

demonstrating the principle of "differentiation and treatment" have important practical significance for development and internationalization of Chinese medicine. Named entity identification has been widely used in many fields, such as: financial field [1], product identification [2], microblog text [3] and military text [4]. In the field of medical text identification, Yang S C et al. [5] identified the symptoms and pathogenesis of the medical case with complicated structural features according to the grammatical features of ancient medical cases in the Ming and Qing Dynasties; Gao J S et al. [6] adopted conditional random field method to extract disease names on web page; Lun W U et al. [7] analysed the effect of conditional random field and maximum entropy Markov model in identification of TCM literature entities; Yang J F et al. [8] combined the characteristics of Chinese electronic medical record and pro-posed labelling system of named entity and entity relationships suitable for Chinese electronic medical records; Feng Y E et al. [9]identified entities such as diseases, clinical symptoms and surgical operations in Chinese electronic medical records; Wang H et al. [10] identified tumor cases based on rule and conditional random field algorithm.

Named entity identification methods fall into two categories: rule and dictionary-based methods and machine learning-based methods. Rule and dictionary-based methods rely heavily on dictionaries and rule bases, and have low identification capabilities for ambiguous words and unregistered words [11]; machine learning-based methods are fast, efficient, and have good portability. Common methods include: Hidden Markov Model (HMM) [12], Maximum Entropy Hidden Markov Model (MEMM) [13], Conditional Random Field Model (CRFs) [14]. After comparison, CRFs perform best in terms of ease of use, stability, and accuracy [15-16], which is superior to HMM and MEMM algorithms in terms of output independence assumptions and unavoidable marker bias problems [17].

In view of vague expression and unclear title in the text of Chinese medical case, based on conditional random field, this paper proposes a named entity identification method based on multi-feature template modification. Using the 12,000 medical cases collected, the author analyzes text features, part-of-speech features and label features of the three types of entities in TCM, namely four diagnostic methods of TCM, syndrome pattern and therapy, defines the template and trains CRFs model, establishes the correlation between text features and named entity categories and lexemes. By extracting text information of four diagnostic methods of TCM- Syndrome-therapy, the author explains the principle of "syndrome differentiation and treatment", to provide a scientific basis for experience inheritance and knowledge acquisition.

2. Methodology

2.1. Conditional Random Field

Conditional random field is defined as follows: Through word segmentation, data cleaning and feature labelling, a text input sequence x ($x=(x_1,x_2,...,x_n)$) is obtained, and the model parameters are obtained through training, and the conditional probability of the required corpus tagging combination y is predicted.

When the input variable is x and the output variable is y, the conditional probability $P(y|x)$ can be defined as Equation (1) and (2).

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \quad (1)$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \quad (2)$$

Where, λ_k is defined as the corresponding weight, t_k and s_l are defined as characteristic function, $Z(x)$ is defined as a normalization factor.

This paper proposes a named entity identification method based on feature template modification, which is used to obtain named entities of four diagnostic methods of TCM, syndrome pattern and therapy, and establish knowledge base of syndrome differentiation and treatment. It mainly includes the following steps.

(1) Extract sentences. First, four diagnostic information, syndrome pattern and therapy sentences are extracted from Chinese medical cases;

(2) Classify sentences based on modifiers. Keep the current sentence and remove the denied and possible sentences;

(3) Word segmentation, data cleaning and typos modification. The text is cut into single words, with all the numbers, units, punctuation, etc. removed; complete data cleaning: for instance, "血淤(blood stagnation)" is modified to "血瘀(blood stagnation)"; non-meaningful words are removed: for instance, "脉滑而数(slippery and rapid pulse)" is modified into "脉滑数(slippery, rapid pulse)" after removal of non-meaningful words;

(4) Corpus tagging is made for each token in the training data set. Identified entity is recognized; entity is identified according to the feature list, the entities here are four diagnostic methods of TCM (ZS), syndrome pattern (ZX) and therapy (ZF);

(5) Text category and lexeme are output to obtain the correlation of four diagnostic methods of TCM- syndrome pattern-therapy.

2.2. Feature Automatic Labelling

Selecting appropriate feature to describe implicit semantic information is the key to entity identification. The term identification features include character (W), part of speech (P), left- right designators (L and R) and Chinese medical case term (Z) as feature pairs for medical case labeling. These features have a good degree of differentiation and are easily labeled automatically.

2.2.1. Character Features (W)

Word-based word segmentation method is used to process medical case text information. For example, "喘促不得卧(hasty panting with inability to lie down)" is divided into "喘/促/不/得/卧/(chuan/chu/ bu/ de/ wo/)".

2.2.2. Part of speech feature (P)

Part of speech is divided into verbs, nouns, adjectives and prepositions. Example of original medical case corpus is as follows: "近几天,患者心悸伴有喘咳、水肿5年,加重10天,10天来患者心悸、喘咳、水肿明显加重,经用镇静强心利尿等西医治疗,逐渐加重。现症见心悸不宁,喘促不得,卧倦怠无力。阳虚水饮。温阳利水泻肺平喘。(In recent days, the patient has palpitation accompanied by cough and edema for 5 years, with exacerbation for 10 days. The palpitation, cough and edema of the patient have been aggravated in the past 10 days. Treatment with Western medicine for calming, strengthening heart and dieresis is not effective, with conditions aggravated. Current symptoms include palpitation and restlessness, hasty panting with inability to lie down, tiredness and weakness, edema due to yang insufficiency. The treatment is warming yang for dieresis, removing heat from lung and relieving asthma.)".The result of part-of-speech tagging is as follows: " 近[jin]/n几[ji]/n天[tian]/n患[huan]/n者[zhe]/n心[xin]/n悸[ji]/v伴[ban]/v有[you]/v喘[chuan]/v咳[ke]/v水[shui]/n肿[zhong]/v加[jia]/v重[zhong]/v患[huan]/v者[zhe]/n心[xin]/v悸[ji]/n

喘[chuan]/v咳[ke]/v水[shui]/n肿[zhong]/v明[ming]/adv显[xian]/adv加[jia]/v重[zhong]/v经[jing]/v用[yong]/v镇[zhen]/v静[jing]/v强[qiang]/v心[xin]/n利[li]/v尿[niao]/v等[deng]/v西[xi]/n医[yi]/n治[zhi]/v疗[liao]/v逐[zhu]/adv渐[jian]/adv加[jia]/v重[zhong]/v现[xian]/adv症[zheng]/n见[jian]/v心[xin]/n悸[ji]/v不[bu]/adv宁[ning]/v喘[chuan]/v促[chu]/v不[bu]/adv得[de]/v卧[wo]/v倦[juan]/v怠[dai]/v无[wu]/adv力[li]/v阳[yang]/n虚[xu]/n水[shui]/n饮[yin]/n温[wen]/n阳[yang]/n利[li]/n水[shui]/n泻[xie]/n肺[fei]/n平[ping]/v喘[chuan]/v".

2.2.3. Left-right designator features (L) and (R)

Named entities such as four diagnostic methods of TCM, syndrome pattern and therapy often appear together with specific predicates, verbs, and adverbs. Some words appear on the left side of a named entity and are called left designators. Those on the right side are referred to as right designators.

In terms of four diagnostic methods of TCM: designator will appear in places adjacent to entities of four diagnostic methods. For example: "头痛,伴有恶心(headache, accompanied by nausea","伴有[accompanied]" here can indicate impending appearance of entities.

In terms of syndrome pattern: designators adjacent to syndrome pattern are often accompanied by "以致[so that]", "之势 [state] ", etc.

In terms of therapy: designators adjacent to therapy are often accompanied by "予[given]", "宜[appropriate]", "仍予[still given]", "治予[treated] " and "治宜[treatment]", etc.

2.2.4. Term features (Y)

Chinese medical case entities include terms describing organs and shapes of human organs, such as "头[head]", "眼[eye]", "舌[tongue]", "火[fire]", and the like to establish a term designator lexicon to facilitate feature identification.

In terms of four diagnostic methods of TCM: for the words describing the substance of the organs and the body's pathological substances, such as "头痛(headaches)" and "眼干(dry eyes)" the first word of these two symptoms is organ system: "头[head]" and "眼[eye]". Another example: for "汗出(sweating)", "尿黄(yellow urine)", the first word of these two symptoms is the body's pathological substance: "汗[sweat]", "尿[urine]".

In terms of syndrome pattern: there are theory of yin-yang and five elements represented by "金[gold]、木[wood]、水[water]、火[fire]、土[earth]", endogenous five evils represented by "风[wind]、寒[cold]、湿[humidity]、燥[dryness]、火[fire]"and description of the mechanism of internal organs of the body, for instance: "扶土抑木(supporting earth to inhibit wood)"、"培土生金(reinforcing earth to generate metal)"、"炽火不降(unyielding fire)"、"寒湿下注(pouring cold dampness)"、"脾虚湿盛(insufficiency of the spleen with overabundance of dampness)"、"心阴不足(insufficiency of heart yin)" and so on.

In terms of therapy: Usually, it is in 4 words or 8 words pattern, such as "疏肝解郁(relieving the depressed liver)", "益气养阴(tonifying qi and yin)"、"活血化瘀(promoting circulation and removing stasis)" and so on.

2.3. Generate Corpus Sequence

Word segmentation and automatic labelling of features generate corpus observation sequences and output feature sequences. "T" indicates that the corpus complies with the labelling feature, "F"

indicates that the corpus does not conform to the labelling feature, and ZS, ZX and ZF represent four diagnostic methods of TCM, syndrome pattern and therapy respectively. Category label of Chinese medical case information is shown in Table 1.

Table 1 Category Label Table of Chinese Medical Case.

Category	Symbol	Examples
four diagnostic methods of TCM	ZS	"心悸不宁 (Palpitation and restlessness)"、"喘促(panting)"、"倦怠(tiredness)"、"畏寒(chilly)"、"肢冷(cold limb)"、"纳差(poor appetite)"、"头晕(dizziness")、"恶心(nausea)"、"口干(thirst)"、"腹胀(abdominal distention)"、"便秘(constipation)"、"尿少(oliguria")
syndrome pattern	ZX	"阳虚(yang deficiency)"、"水饮(excessive fluid)"
therapy	ZF	"温阳利水"(Warming yang to promote diuresis)、"泻肺平喘"(removing heat from lung and relieving asthma)

"BIO" method is adopted for labeling. "B" represents the first character of the entity, "I" represents the non-initial character of the entity, "O" represents non-entity character. For instance, for the entity of four diagnostic method of "近几天心悸(palpitation and restlessness in recent days)", "近[recent]"、"几[several]"、"天[day]" are non-entity characters and thus labeled as "O", while"心[heart]"、"悸[palpitation]" are respectively set "ZS-B" and "ZS-I". The category label and annotation of the example are shown in Table 2.

Table 2 "BIO" Method for Labelling.

Observation Sequence					Output Sequence Position and Category
Chinese Character (W)	Position and Category	Left Designator (L)	Right Designator (R)	Term (Y)	
"近[jin]"	n	F	F	F	O
"几[ji]"	n	F	F	F	O
"天[tian]"	n	F	F	F	O
"患[huan]"	n	F	F	F	O
"者[zhe]"	n	F	F	F	O
"心[xin]"	n	F	F	T	ZS-B
"悸[ji]"	v	F	F	T	ZS-I
"伴[ban]"	v	T	F	F	O
"有[you]"	v	T	F	F	O
"喘[chuan]"	v	F	F	T	ZS-B
"咳[ke]"	v	F	F	T	ZS-I
"阳[yang]"	n	F	F	T	ZX-B
"虚[xu]"	v	F	F	T	ZX-I
"水[shui]"	n	F	F	T	ZX-B
"饮[yin]"	v	F	F	T	ZX-I
"温[wen]"	v	F	F	T	ZF-B
"阳[yang]"	n	F	F	T	ZF-I
"利[li]"	v	F	F	T	ZF-I
"水[shui]"	n	F	F	T	ZF-I

3. Results and Discussion

3.1. Evaluation Standard

The indicators extracted for information evaluation include: accuracy rate (P), recall rate (R) and F-measure (F), which are defined as Equation (3)-(5).

$$\text{Precision(P)}=\frac{\text{the correct amount of extracting}}{\text{the actual amount of extracting}} \tag{3}$$

$$\text{Recall rate(R)}=\frac{\text{the amount of correct extracting}}{\text{the amount of proper extracting}} \tag{4}$$

$$\text{F measure value(F)}=\frac{2P\times R}{P+R} \tag{5}$$

3.2. Evaluation Standard

3.2.1. Feature identification

In this paper, context features are extracted from context window [-2, 2] with window size 5, and the feature space is called "5-word window" [18-19]. According to structural characteristics of the template, the common features are set to 13 categories. A set of feature templates are represented as "letter + number". Where, "W" denotes the word itself, "P" denotes part of speech, "L" and "R" denote left and right indicative connectives, and "Y" denotes TCM term features. According to the context, 19 common feature identifiers are set, and Table 3 shows the common feature identifier and meaning.

Table 3 Identification and Meaning.

Number	Identification	Meaning	Number	Identification	Meaning
1	W_-2	first second Chinese character	2	W_-1	first Chinese character
3	W_0	current Chinese character	4	W_1	latter Chinese character
5	W_2	latter second Chinese character	6	P_-2	part of speech of the first second Chinese character
7	P_-1	part of speech of the first Chinese character	8	P_0	part of speech of the current Chinese character
9	P_1	part of speech of the latter Chinese character	10	P_2	part of speech of the latter second Chinese character
11	L_-2	left designator of the first second Chinese character	12	L_-1	left designator of the first Chinese character
13	R_1	right designator of the latter Chinese character	14	R_2	right designator the latter second Chinese character
15	Y_-2	term of the first second Chinese character	16	Y_-1	term of the first Chinese character
17	Y_0	term of the current Chinese character	18	Y_1	term of the latter Chinese character
19	Y_2	term of the latter second Chinese character			

3.2.2. Experimental design

Tmpt_1, Tmpt_2, Tmpt_3, and Tmpt_4 were used as templates to complete three groups of experiments for testing. The influence of feature selection and window size on the identification effect was tested. The template is defined and shown in Table 4. The experimental design is shown in Table 5.

Table 4 Template Definition.

Template name	Template definition
Tmpt_1	W_-1,W_0,W_1,W_-1/W_0,W_0/W_1,P_-1,P_0,P_1,P_-1/P_0,P_0/P_1
Tmpt_2	W_-2,W_-1,W_0,W_1,W_2,W_-1/W_0,W_0/W_1,W_-2/W_0, W_0/W_2, P _-2,P_-1,P_0,P_1,P_2,P_-1/P_0,P_0/P_1, P_-2/P_0,P_0/P_2
Tmpt_3	W_-2,W_-1,W_0,W_1,W_2,W_-1/W_0,W_0/W_1,W_-2/W_0, W_0/W_2, P _-2,P_-1,P_0,P_1,P_2,P_-1/P_0,P_0/P_1, P_-2/P_0,P_0/P_2,L_-2/W_0,L_-1/W_0, W_0/R_1,W_0/R_2
Tmpt_4	W_-2,W_-1,W_0,W_1,W_2,W_-1/W_0,W_0/W_1,W_-2/W_0, W_0/W_2, P _-2,P_-1,P_0,P_1,P_2,P_-1/P_0,P_0/P_1, P_-2/P_0,P_0/P_2,L_-2/W_0,L_-1/W_0, W_0/R_1,W_0/R_2, Y_-2/W_0,Y_-1/W_0, W_0/Y_0,W_0/Y_1,W_0/Y_2

Table 5 Experimental design

Experiment Name	Template Selection	Experiment Purpose	Description
Experiment 1	Tmpt_1 Tmpt_2	Analyze the effect of context window size on identification of named entities when set to 3 and 5 respectively	In this set of experiments, the features only include character feature label and part-of-speech feature label; Tmpt_1 context window size is set to 3, Tmpt_2 context window size is set to 5.
Experiment 2	Tmpt_2 Tmpt_3	In the case of window size set to 5, the effect of addition of left and right feature label on named entity identification.	Tmpt_3 context window size of is set to 5, including the left and right identification feature labels in addition to character feature label and part-of-speech feature label.
Experiment 3	Tmpt_3 Tmpt_4	The purpose of this group of experiments is to analyze the effect of addition of term feature label on identification of named entities when the window size is set to 5.	Tmpt_4 context window size is set to 5, including character features, part of speech features, left and right designators and term labels

3.3. Analysis of Results

3.3.1. Analysis of Experiment 1

Tmpt_1 and Tmpt_2 were used as experimental templates and set to 3 and 5 respectively. Table 6 shows increase and decrease in experimental results when the context window is set to 5 as compared with context window height set to 3.

Table 6. Effect of window change on the result

Category Name	P(%)	R(%)	F(%)
four diagnostic methods of TCM	+0.45	+0.47	+0.46
syndrome pattern	+0.04	+0.38	+0.14
therapy	+1.83	+1.12	+1.28

Experiments have found that when the context window changes, the accuracy will change. It is reasonable to use a context window with a length of 5 for medical entity identification. The effect of different types of named entities differs. The average character size of the three types of entities is: 3.17 characters for four diagnostic methods of TCM, 2.21 characters for syndrome patterns and 4.78 characters for therapy. Seen from the improvement of the effect: F value of the four diagnostic methods of TCM increases by 0.46%, that of syndrome pattern increases by 0.14%, and that of therapy increases by 1.28%. This shows that identification effect is fine when the entity identification effect is similar to the selected context window length.

3.3.2. Analysis of experiment 2

Tmpt_2 and Tmpt_3 were used as experimental templates. After left and right identifiers are added to template Tmpt_3, identification effect is obvious. Where, improvement in therapy effect is the most obvious. However, because of particularity of TCM terminology, some terms cannot be well identified, which affects the experimental results. The effect of feature selection category identifier on results is shown in Table 7.

Table 7. Effect of adding left and right identifiers on the result

Category Name	P(%)	R(%)	F(%)
four diagnostic methods of TCM	+7.17	+6.23	+0.19
syndrome pattern	+5.37	+5.48	+0.42
therapy	+5.86	+4.76	+0.84

3.3.3. Analysis of Experiment 3

In Experiment 3, a new experimental group was added. The template was Tmpt_4 which was compared with Tmpt_3 to see the effect of addition of term feature identifier on named entity identification. Among the three experimental groups, the new experimental group shows the best effect, shown in Table 8.

Table 8. Optimal recognition results

Template Name	Category Name	P(%)	R(%)	F(%)
Tmpt_1	four diagnostic methods of TCM	67.39	60.67	63.85
	syndrome pattern	72.18	68.92	70.51
	therapy	78.26	71.56	74.76
	average value	72.61	67.05	69.72
Tmpt_2	four diagnostic methods of TCM	67.84	61.14	64.32
	syndrome pattern	72.22	69.30	70.73
	therapy	80.09	72.68	76.21
	average value	73.38	67.71	70.43
Tmpt_3	four diagnostic methods of TCM	75.01	67.37	70.99
	syndrome pattern	77.59	74.78	76.16
	therapy	85.95	77.44	81.47
	average value	79.52	73.12	76.19
Tmpt_4	four diagnostic methods of TCM	90.19	89.78	89.98
	syndrome pattern	91.03	91.53	91.28
	therapy	90.56	90.03	90.29
	average value	90.68	90.45	90.56

By comparing F-values of various named entities in Table 8, it is found that the best template is Tmpt_4 whose accuracy, recall and F-measure averages are: 90.68%, 90.45% and 90.56%, respectively, with great improvement in identification performance. This suggests that a rich feature set can improve the accuracy of named entity identification.

There are still a lot of gaps between the experimental results and entity identification results in other fields, which are due to the influence of Chinese medical case grammar and terminology features. For example, in Chinese medical case, the word "下[xia] " often appears: "发汗吐下后,虚烦不得眠(after sweating and vomiting, one can't sleep with dysphoria)", "寸口脉浮大,而医反下之(cunkou pulse floats greatly, which is treated by purgation)", "脉浮而大,心下反硬(pulse floats greatly, epigastric fullness and rigidity)". In the first two, the word "下[xia]"refers to "下法[purgation]" in TCM treatment, while in the last one, it only represents position. Words like these can affect accuracy of term identification. In view of this situation, it is necessary to further modify identification system with the help of dictionaries and rules.

3.3.4. Comparison with the existing methods

After consulting the literature, Feng Lizhi [20] proposed a hybrid method based on Bootstrapping for the clinical medical corpus of TCM, with F value reaching 87%; Yuan Yuhu [21] carried out named entity extraction experiment on symptom terms with CRFs model, and optimal F value of open test reached 87%; in this paper, average F value of named entity identification in the medical case is higher than the two, reaching 90.51%.

4. Conclusions

The diversity of Chinese medical case texts and the characteristics of description complexity determine the difficulty in identifying named entities of Chinese medical case texts. This paper applied a conditional random field method to propose a named entity identification method for Chinese medical cases based on multi-feature template modification. Analysis and treatment was carried out for 12,000 pieces of medical case information of heart disease. Labelling methods of character features, part of speech features, left and right designator features and term features were proposed in combination with the characteristics of Chinese medical case texts. CRFs models were trained using the labelled data to identify four diagnostic methods of TCM, syndrome pattern and therapy entities. Through experimental verification, accuracy rate, recall rate and F-measure have been greatly improved after adding left and right designator feature and term feature identifiers. Through continuous accumulation of medical cases and more reasonable parameter settings, as well as further rational setting of eigenvalues, the named entity identification method provides more valuable references and evidence for building triple correspondence of four diagnostic methods of TCM-syndrome pattern-therapy and making scientific argumentation of syndrome differentiation and treatment.

Acknowledgements

This paper is financed by the National Social Science Foundation of China (16BGL181).

References

[1] Park G, Kim H. (2018) Low-Cost Implementation of a Named Entity Recognition System for Voice-Activated Human-Appliance Interfaces in a Smart Home. Sustainability, 10(2).

[2] Ekbal A, Saha S.(2016) Simultaneous feature and parameter selection using multiobjective optimization: application to named entity recognition. International Journal of Machine Learning & Cybernetics, 7(4):597-611.

[3] Wang H L.(2017) Named entity recognition of Chinese microblog product based on word-vector clustering. Journal of Lanzhou University of Technology.
[4] Feng Y T, Zhang H J, Hao W N. (2015) Named Entity Recognition for Military Text. Computer Science.
[5] Yang S C, Yue J I, Zhao L P.(2017) Study of Ancient Chinese Word Segmentation Based on Conditional Random Field[J]. Computer Knowledge & Technology.
[6] Gao J S, Ying-Ying L I, Liu L, et al. (2016) Research on Construction of the Knowledge Discovery Model Based on Linked Data. Information Science.
[7] Lun W U, Lei L, Haoran L I, et al. (2017) A Chinese Toponym Recognition Method Based on Conditional Random Field. Geomatics & Information Science of Wuhan University, 42(2):150-156.
[8] Yang J F, Qiu-Bin Y U, Guan Y, et al. (2014) An Overview of Research on Electronic Medical Record Oriented Named Entity Recognition and Entity Relation Extraction. Acta Automatica Sinica, 40(8):1537-1562.
[9] Feng Y E. (2011) Intelligent Recognition of Named Entity in Electronic Medical Records. Chinese Journal of Biomedical Engineering, 30(2):256-262.
[10] Wang H, Zhang W, Zeng Q, et al. (2014) Extracting important information from Chinese Operation Notes with natural language processing methods. Journal of Biomedical Informatics, 48(C):130-136.
[11] Ekbal A, Saha S, Sikdar U K. (2016) on active annotation for named entity recognition. International Journal of Machine Learning & Cybernetics, 7(4):623-640.
[12] Shruthi S, Jiljo, Pranav P V. (2016) A study on named entity recognition for malayalam language using tnt tagger & maximum entropy markov model. International Journal of Applied Engineering Re-search, 11(8):5425-5429.
[13] Gassiat E, Rousseau J. (2016) Nonparametric finite translation hidden Markov models and extensions. Bernoulli, 22(1):193-212.
[14] Bourgeauchavez L L, Endres M S L, Powell M R, et al. (2016) Mapping Boreal Peatland Ecosystem Types from Multi-Temporal Radar and. Canadian Journal of Forest Research, 47(4).
[15] Guo H. (2015) Accelerated Continuous Conditional Random Fields for load forecasting. IEEE Trans-actions on Knowledge & Data Engineering, 27(8):2023-2033.
[16] Kosov S, Shirahama K, Li C, et al. (2017) Environmental Microorganism Classification Using Conditional Random Fields and Deep Convolutional Neural Networks. Pattern Recognition, 77:248-261.
[17] Zhu Y, Liu J, Yeqiang X U, et al. (2016) Chinese word segmentation research based on Conditional Random Field. Computer Engineering & Applications.
[18] Jiang T. (2016) Research on Chinese Automatic Terminology Extraction Based on SVR Model. Information Studies Theory & Application.
[19] Wang S. (2016) Analyzing left branch extraction in Chinese noun phrases under phase theory. Modern Foreign Languages.
[20] Liu K, Zhou X Z, Jian Y U, et al. (2014) Named Entity Extraction of Traditional Chinese Medicine Medical Records Based on Conditional Random Field. Computer Engineering.
[21] Eftimov T, Seljak B K, Korošec P. (2017) A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. Plos One, 12(6).