Journal of Computers Vol. 28, No. 5, 2017, pp. 210-219
doi:10.3966/199115992017102805019

Closeness Based New Word Detection Method for Mechanical Design and Manufacturing Area

Qiuyuan Chen[12], Guang Cheng[13*], Di Li[14] and Jian Zhang[14]

[1] Beijing Engineering Research Center of Smart Mechanical Innovation Design Service, Beijing 100020, P. R. China

[2] Computer Science and Technology, Information College, Beijing Union University
chenqyld@163.com

[3] Mechanical and Electrical Engineering, Beijing Union University
chengguang@buu.edu.cn

[4] Manufacturing Information, Mechanical and Electrical Engineering, Beijing Union University
915410987@qq.com

Received 25 October 2016; Revised 13 June 2017; Accepted 26 June 2017

Abstract. Named entity recognition has been widely used in the area of information retrieval, but the common methods cannot accurately identify the proper nouns from a particular domain. In order to solve the named entity recognition for mechanical design and manufacturing area, this paper proposes a closeness based method, in order to identify the proper nouns. First, we calculate the entropy about each character, and then define the closeness between two adjacent words based on the lexical features and statistic features. Finally we use the logistic regression algorithm to determine the weights in the closeness definition. The proposed method can recognize proper nouns more accurately and efficiently for mechanical design and manufacturing area.

Keywords: information entropy, left(right) entropy, logistic regression, mechanical design, named entity recognition, new word detection

1 Introduction

In the recent years, the search engine and network news become more and more popular in the internet applications. In order to make the information retrieval more convenient in the field of mechanical design and manufacturing, the difference between this area and the common area must be studied. Named entity recognition is the foundation of information retrieval [1] and it determines the quality of the search engine. Therefore, we focus on the named entity recognition in the mechanical design and manufacturing area. Previous studies show that about 60% of segmentation errors are caused by new words, so effectively identify new word is an urgent problem in the field of Chinese lexical analysis [2-7].

Named Entity (Named Entity, NE) is the basic unit of information in the text. It may be the text proper names, abbreviations or some uniquely identifies, and it is very important to the understanding of the whole text. Named entity recognition [15] is the foundation work to the information extraction, question answering system, parsing, machine translation, metadata tagging, and plays an important role in the development of natural language processing technology. The representative methods of named entity recognition are mainly divided into two kinds [8]: rule based method and statistic based method. In the recent years, some methods have been proposed based on a combination of rules and statistics. Zhang, Liu, Zhang and Cheng [3] introduced the concept of role annotation. First, it use the Bayesian algorithm and Viterbi algorithm to automatically annotate the roles, and then use pattern matching method to obtain

* Corresponding Author

a new set of candidate words, finally it uses the filter rules to obtain of the new words. He Min et al uses the context analysis of adjacency, position analysis and degree of coupling to detect the new words [9]. Shi et al. proposed a new word detection method based on mutual information, frequency ratio and likelihood ratio [10]. Lin et al. proposed a new method, in which the new words are filtered by the combination of the mutual information, the location probability of the word and the internal pattern [11]. Although the statistics-based approaches recognize some new words, it also generates a lot of garbage words. In the other words, statistical based methods can improve the recall ratio at the cost of reducing the accuracy. Therefore, Huo Shuai et al. proposed an iterative algorithm for context entropy get the candidate word list, and then use the combination of lexical features and statistical features to filter them [12].

Currently, the common named entity recognition methods cannot effectively identify the proper nouns in the area of mechanical design and manufacturing. Take the heat treatment as an example; it is common to split it into two words, which is heat and treatment. Therefore, the study for the named entity recognition in a special area is important. In this paper, we propose a closeness based method, in order to identify the proper nouns. First, we calculate the left and right entropy about each character, and then define the closeness between two adjacent words based on the lexical features and statistic features.

2 Background

In this part, two important concepts are introduced first since they will constitute two features in the new proposed method.

2.1 The Concept of Information Entropy

For the natural language processing, left (right) entropy is an important statistical characteristic for the string. Left (right) entropy is used to measure the level of active for a word in the context. If the left (right) entropy of a word is high, it indicates that it can form a phrase with more different words. In other words, the left (right) entropy represents the number of different words that are adjacent to it in the context. The technology of the entropy is widely used in the term extraction and the new word detection [14].

An example of the information entropy application is shown in Table 1 and Table 2. Variable X stands for the event of the game between A and B. In the two different case, the differences of probability distribution lead to different entropy result. Generally speaking, if the results are more randomly, the entropy is high.

Table 1. Example 1

Variable X	A win	B win
Probabilit Y	0.9	0.1
Information entropy H	0.325	

Table 2. Example 2

Variable X	A win	B win
Probabilit Y	0.5	0.5
Information entropy H	0.693	

The higher uncertainty of variables, the greater the entropy is, since it needs more information to describe it. For a steady system, the information entropy is low, while for a complicated system, the information entropy is high. Therefore, entropy is said to be a measure of the degree of system complexity. [13].

In order to describe the information of a signal, we should consider all its states and their corresponding probabilities. If the signal has n values: $U_1 \cdots U_i \cdots U_n$, corresponding probability: $P_1 \cdots P_i \cdots P_n$, the occurrence of a value of the signal is independent of each other. Then, the uncertainty of the signal should be the statistical mean of each probability $-logP_i$, and it is referred to as entropy.

[13].

$$H(U) = E[-\log p_i] = -\sum_{i=1}^{n} p_i \log p_i \quad (1)$$

2.2 Hopysis Testing and Likelihood Ratio

A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

Here we use likelihood ratio to find the collocations. In statistics, a likelihood ratio test is a statistical test used to compare the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. In this paper, one hypothesis is used to model the two words are independent with each other, while the other hypothesis is used to model the relationship of co-occurrence. Therefore the ratio can be used to represent the possibility of a new collocation.

3 Closeness Based New Word Detection Method

In this paper, we propose a new method to detect the new word for mechanical design and manufacturing area. The new method is a machine learning method and it contains three important aspects, (1) methodology, (2) features and (3) data set.

For the methodology, we use the logistic regression algorithm to apply on the trained data and get the weighs of each weight. Using this method, we can get the final probability result easily and the learning process is also very efficient.

For the features, five features from the lexical and statistic views are extracted to represent the relationship. In the following part, we first introduce two complex features, entropy and likelihood ratio, and then describe all the feature together. Finally, we propose the machine learning method.

For the data set, we randomly select the words based on the frequency, in order to make sure the sample can contain enough information of the population. For the selected words, manually annotation is used to determine whether they are new words or not. Then logistic regression method is used to find the weights of each feature based on the annotated words.

3.1 Left (Right) Entropy Calculation

Based on the concept of entropy, we can use it to solve new word detection problems in specific areas, since the left (right) entropy can be used to evaluate whether the adjacent two word is a phrase. The formula for the left (right) entropy is as follow [13]:

$$LE = -\sum_{i=0}^{n} p_i(xAB \mid AB)\log_2 p(xAB \mid AB) \quad (2)$$

where x is any word that appears on the left AB in the context

$$RE = -\sum_{i=0}^{n} P_i(ABy \mid AB)\log P(ABy \mid AB) \quad (3)$$

where y is any word that appears on the right of AB in the context.

The left (right) entropy of a word is mainly used to measure the diversity of adjacent word that appear on its left or right side. By calculating the probability of the occurrence for words on the left and right side, the left (right) entropy can be obtained.

The process for calculating the left (right) entropy is as follow:

Step 1. Build a table, referred to as Table, to record the occurrence of each word. By traversing the whole text, each word is noted down and the number of occurrence is obtained.
Step 2. Build the second table, referred to as Table1. For each word w_i, record the number of occurrence of the word on its left or right side $T_i,(v_i)$.
Step 3. Calculate the probability of $T_i,(v_i)$ appearing on the left and right side of w_i.
Step 4. According to the formula , calculate the left entropy and right entropy.

3.2 Likelihood Ratio

The likelihood ratio test is a statistical test used to compare the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model). It simply a number that tells us how much more likely one hypothesis is than the other. Here, the null hypothesis is the two words are independent with each other while the alternative hypothesis is that the two word is dependent with each other. The advantage of likelihood ratios is that they have a clear intuitive interpretation. If the number is large, it means that the two words are more likely to be two independent words.

We list the two hypothesis as following:

Hypothesis 1: $P\left(w^1 \mid w^2\right) = P\left(w^1 \mid \wedge w^2\right)$

Hypothesis 2: $P\left(w^1 \mid w^2\right) != P\left(w^1 \mid \wedge w^2\right)$

In order to describe the different cases.we define the variable p, p_1, p_2 .The number of occurrenc- es of the word (w^1, w^2, w^{12}) in the document are referred to as (c, c_1, c_2).We can get

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1} \tag{4}$$

c_1 : the number of times A appears

c_2 : the number of times B appears

c_{12} : the number of times AB appears

Given binomial distribution

$$b\left(k; n, x\right) = \binom{n}{k} X^k \left(1 - x\right)^{(n-k)} \tag{5}$$

The number of occurrence times for w^1, w^2, w^{12} can be obtained from the traversing the text. The following two cases represent the probability of event w^2 happening c_2 times in the total N times under hypothesis 1 and hypothesis 2,

$$L(H_1) = b\left(c_{12}, c_1, p\right) b\left(c_2 - c_{12}, N - c_1, p\right) \tag{6}$$

$$L(H_2) = b\left(c_{12}, c_1, p_1\right) b\left(c_2 - c_{12}, N - c_1, p_2\right) \tag{7}$$

From the reference, the formula of likelihood ratio [16] is as follows:

$$\log l = \log \frac{L(H_1)}{L(H_2)} = \log \frac{b(c_{12}, c_1, p)\, b(c_2 - c_{12}, N - c_1, P)}{b(c_{12}, c_1, p_1)(c_2 - c_{12}, N - c_1, P)} \tag{8}$$

In order to apply the theory to the new words detection , we make the following assumptions: the variable a represents the number of A and B occurring simultaneously, and b represents the number of cases that only A happened, while c represents the number of cases that only B happened. The variable d represents the number of the cases that neither A nor B happened. Then we can get the following equations.

$$P = \frac{a + b}{N} \tag{9}$$

$$P_1 = \frac{a}{a+c} \tag{10}$$

$$p_2 = \frac{b}{b+d} \tag{11}$$

For a binomial distribution, the probability density function is as:

$$b(k;n,x) = \binom{n}{k} x^k (1-x)^{(n-k)} \tag{12}$$

In order to simplify the equations, let us note

$$L(k,n,x) = x^k (1-x)^{n-k} \tag{13}$$

With the equations (9)(10)(11) and formula (8), we can get:

$$\log L\left(a, a+c, \frac{a+b}{N}\right) + \log L\left(b, b+d, \frac{a+b}{N}\right) - \log L\left(a, a+c, \frac{a}{a+c}\right) - \log L\left(b, b+d, \frac{b}{b+d}\right) \tag{14}$$

According to the formula (13) and (14), The following results can be obtained. Make A

$$\log L = \log\left(\frac{(a+b)*(a+c)}{N*a}\right)^a * \log\left(\frac{(c+d)*(a+c)}{N*c}\right)^c * \log\left(\frac{(a+b)*(b+d)}{N*b}\right)^b * \log\left(\frac{(c+d)*(b+d)}{N*c}\right)^d \tag{15}$$

From the reference, likelihood ratio formula should be -2A, [18] we can get:

$$\begin{aligned} LLR(x,y) = 2(a*\log\frac{a*N}{(a+b)*(a+c)} + b*\log\frac{b*N}{(a+b)*(b+d)} \\ + c*\log\frac{c*N}{(c+d)*(a+c)} + d*\log\frac{d*N}{(c+d)*(b+d)}) \end{aligned} \tag{16}$$

The result will be a feature to indicate whether the two words are independent with each other.

3.3 Calculation Method of Closeness

We will introduce the concept of closeness as a measure to detect the new words. The closeness between A and B is defined to describe the probability that A and B can be combined to a new phrase. The result of the closeness is closely related to the following five features. Using the logic regression method, the weight for each feature can be determined when calculating the closeness property. Based on the results of the closeness, we can detect new words.

Adjacent probability. The probability of two words that is adjacent to each other. It can be calculated as (The number of times they are adjacent to each other)/(total times they appear).

Similar probability. The probability that two words appear at the same time within a fix-length string. It can be calculated as: (the number of times appearing at the same time)/ (total times they appear).

Monotonic probability. The probability that two words appear in the reverse order. It can be calculated as: (number of reverse appearing) / (total times they appear).

Left (right) entropy. Using the following equations to calculate:

$$LE = -\sum_{i=0}^{n} p_i(xAB \mid AB) \log_2 p(xAB \mid AB) \tag{17}$$

$$RE = -\sum_{i=0}^{n} p_i(AB\,y \mid AB) \log_2 p(AB\,y \mid AB) \tag{18}$$

The likelihood ratio. Firstly, list the assumptions as in Table 3.

Table 3. Assumed definition

	Times for A appears	Times that A doesn't appear
Times for B appears	a	b
Times for B doesn't appear	c	d

The definition for the likelihood ratio is as follow ($N = a + b + c + d$):

$$\begin{aligned} \mathrm{LLR}(x,y) = 2(&a^* \log \frac{a^* N}{(a+b)^*(a+c)} + b^* \log \frac{b^* N}{(a+b)^*(b+d)} \\ &+ c^* \log \frac{c^* N}{(c+d)^*(a+c)} + d^* \log \frac{d^* N}{(c+d)^*(b+d)}) \end{aligned} \tag{19}$$

In summary, the closeness depends on the adjacent probability, similar probability, monotonic probability, left (right) entropy and likelihood ratio. Adjacent probability and similar probability are focus on the probability that two words appear at the same time. Monotonic probability pays attention to the orders of the words, since the words of a phrase usually appear in a fix sequence. High left (right) entropy means that the word on its left (right) side are more likely to make up a new phrase with it. Likelihood ratio can measure the probability of a word from the perspective of hypothesis testing. Assume the weights for the five features are w_1, w_2, w_3, w_4, w_5, then closeness

$$= w_1 * f_1 + w_2 * f_2 + w_3 * f_3 + w_4 * f_4 + w_5 * f_5 \tag{20}$$

Where f_i represents the values of the features. Based on the concept of closeness, we propose a new word recognition method, including the following three steps: (1) preprocessing and word segmentation for the given text; (2) calculate the five features for the segmented results; (3) based on the results of manual annotation, use the logistic regression to calculate the weights for the five features.

In the second step, we need to calculate the values of five features for each word. Adjacent probability, similar probability and monotonic probability can be obtained by traversing the whole text once. In the process of traversal, the number of times that corresponding relationship occurs is recorded, so the computational complexity is linear with the length of the text. The feature value for left (right) entropy is calculated by the method 3.1. The process of calculating likelihood ratio is similar with left (right) entropy. For any pair of words A and B, in the process of calculating the left (right) entropy, we have figured out the times that both A and B appear at the same time. Then, the times that A appears and B does not appear can be obtained by the times of A appears subtracting the times of AB appears. Similarly, other parameters of likelihood ratio can be calculated. In the third step, use the logistic regression method to determine the weight of each feature based on the manual annotation results. For a manual annotation case, $(f_1, f_2, f_3, f_4, f_5, R)$, where R can be 0 or 1, indicating whether it is a new word. Logistic regression uses the annotation result as the training set to generate the weights for the five features. Based on the result of the weights, the result R can be re-calculated for each annotation case.

3.4 Thresthold of the New Word

Based on the result of weights, the closeness between any words can be calculated with the feature values. How to find the threshold is critical for the new word recognition. Here we use the cross-verification method to find the threshold.

For the data set, we divide the data set into three parts, training set, verification set and testing set. All the data are divided into three parts randomly. The result of the training set is manually annotated and then decision tree method is used to find the estimated threshold. The learned result is verified using the verification data set and adjusts the threshold. The accurate of the selected threshold can be tested using the test set.

The overall flow for the closeness based new word detection method is as Fig. 1.

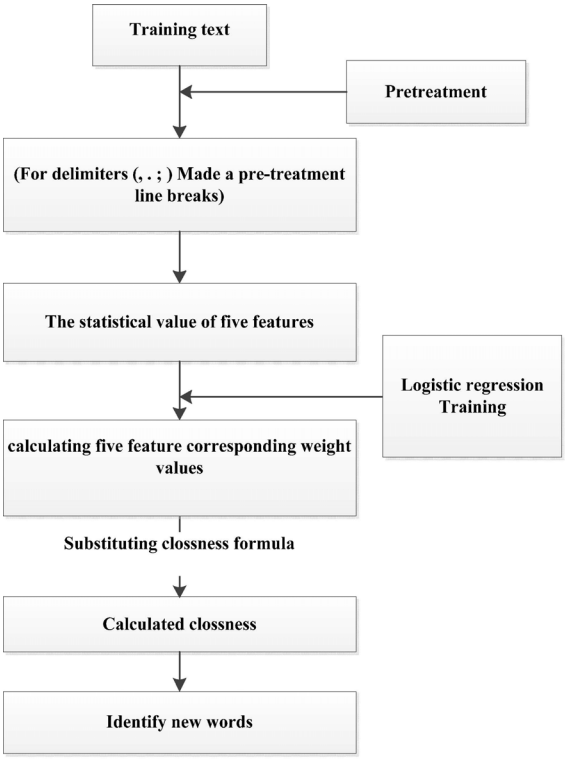

Fig. 1. Experimental flow chart

Analysis of experimental results. For one article, we should pay attention to the keyword, such as nouns and verbs. Therefore, deleting the negligible words is an necessary and useful step. The preprocessing step neither affects the meaning of whole context, nor reduces the information. In this step, the commonly used separators, punctuation, and auxiliary words are filtered out. The flow of the experiment is as Fig. 2.

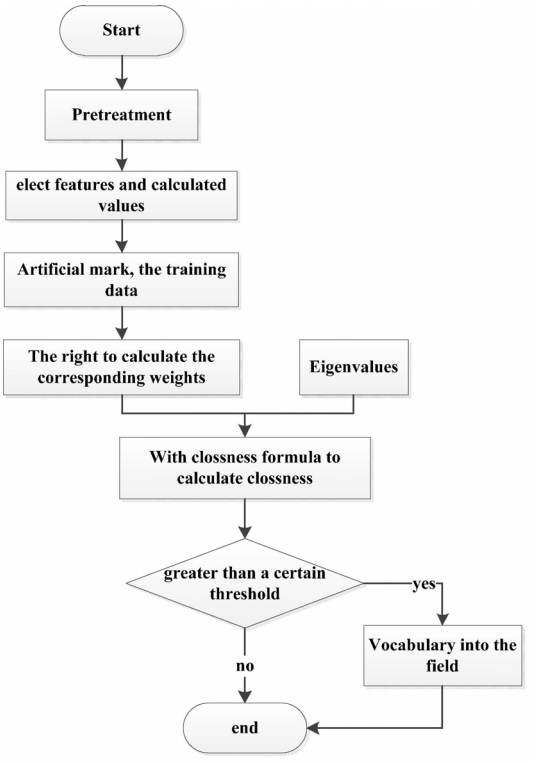

Fig. 2. Experimental flow chart

Training process:

(1) Calculate features: for the words in the candidate set, calculate all their feature values.

(2) Manual annotation: determining whether the two words should be a phrase. 1 stands for yes while 0 stands for no. In this paper, the fixed-length parameter for calculating the similar probability feature is 5. When calculating the feature for left (right) entropy, left entropy and right entropy are considered as two features. We use the logistic regression to train the test data, and obtain the weight for each feature in the closeness definition. The flow for the experiment is shown in the Fig. 1. First, use a web crawler tool to obtain the whole text from the site, and then process the special characters and word segmentation, next, calculating all the features for each candidates, finally, logistic regression algorithm is used to get the weight results.

For the data set, we divide the data set into three parts, training set, verification set and testing set. All the data are divided into three parts randomly. The result of the training set is manually annotated and then a initial value is determined. Using the verification data set we can continuely adjust the threshold and find the value with high confident level (95%).

In the experiment, more than 300 web pages are processed, and there are 195 new words. The proposed method detect 156 new words from the test set. Among these detected new words, 147 words are really new ones. So the accuracy rate is 94.42% while the recall rate is 75.38%. Then we use the tool ICTCLAS [17] to process the 300 pages to get the comparison results. It can identify 80 new words, while 75 of them are really new words, so the accuracy rate is 93.75%, while the recall rate is 41.02%. The comparison result is shown in Table 4.

Table 4. Comparison Tablle

Method	Precision	Recall	F-measure
This method	94.42%	75.38%	83.89%
NLPIR	93.75%	41.02%	57.07%

In order to better illustrate the results, we list some detailed result in Table 5. The first column lists the new words. The second column lists the closeness value for the given word using the proposed method. The third column gives the results from the method ICTCLAS. The last columns gives the comparison results. From the experimental results, we can see that the proposed method can more effectively than the method ICTCLAS in new word detection for mechanical design and manufacturing area.

Table 5. RESULT

New word	clossness	Ictclas new word recognitiond	Comparison results
Shell	type [1.000000] hull	Shell	Both can be identified
Investment Casting	fuse [0.729412] mold [0.729412] cast metal	Investment Casting	Both can be identified
Expert System	expert [0.729412] system	Expert System	Both can be identified
Fusible Pattern	melt [0.733333] mold	Fusible Pattern	Both can be identified
Stone-form Fracture	stone [0.776471] shape [0.776471] Fracture	Stone-form Fracture	Both can be identified
Superally	temperature [0.729412] alloy	Superally	Both can be identified
Grommet	protect [1.000000] ring	Grommet	Both can be identified
gray iron	Gray [0.917647] cast iron	(Gray)(cast iron)	Only this method can be identified
Screw Press	Stone [0.776471] shape [0.776471] Fracture	Stone-form Fracture	Only this method can be identified
Solution Treatment	solidify [0.917647] dissolve [0.917647] process	(solidify)(dissolve)(process)	Only this method can be identified
Huai cleavage fracture	Huai [0.776471] cleavage [0.776471] fracture	(Huai) (cleavage) (fracture)	Only this method can be identified
Low Tempering	hypothermia [0.917647] tempering	(hypothermia)(tempering)	Only this method can be identified
Fire Consumed fracture	fire [1.000000] consume	(fire)(consume)	Only this method can be identified

Table 5. RESULT

New word	clossness	Ictclas new word recognitiond	Comparison results
Precision forgings	accuracy [0.780392] forging	(accuracy)(forging)	Only this method can be identified
Cold shuts	cold [0.917647] partition	(cold)(partition)	Only this method can be identified
Casting Technology	mold [0.917647]technology	(nodular)(cast iron)	Only this method can be identified
Ductile Iron	nodular [0.729412] cast iron	(nodular)(cast iron) (cold)(partition)	Only this method can be identified
Thermal decomposition	heat [0.917647] decomposition	(heat)(decomposition)	Only this method can be identified
Aluminum mold	aluminum [0.90144] mold	(aluminum)(mold)	Only this method can be identified
Stretching mandrel	dabber [0.776471] pull [0.776471] long	(dabber)(pull)(long)	Only this method can be identified
naphthalene fracture	naphthalene [0.776471] shape [0.776471] port)	(naphthalene)(shape)(port)	Only this method can be identified

4 Conclusion

This paper proposed a new named entity recognition method based on the conception of closeness. The closeness based method considered both lexical features and statistical features to find the new word more efficiently. Experimental results show that the new method can detect the new words in the mechanical design and manufacturing area more accurately and efficiently.

References

[1] S. Li, Research on promoting the integration of new Industrialization and informatization, [dissertation] Shenyang: Liaoning University, 2004.

[2] R. Sproat, T. Emerson, The first international Chinese word segmentatio bake off [EB/OL]. <http://acl. ldc.upenn.edu/W/W-03/W03-1719.pdf>, 2013 (accessed 13.03.10)

[3] K. Zhang, Q. Liu, H. Zhang, X.-Q. Cheng, Automatic recognition of Chinese unknown words based on roles tagging, in: Proc. Sighan Workshop on Chinese Language Processing Association for Computational Linguistics, 2002.

[4] Li, Hongqiao, C.-N. Huang, J. Gao, X. Fan, The use of SVM for Chinese new word identification, Lecture Notes in Computer Science 3248(2004) 723-732.

[5] K.J. Chen, W.Y. Ma, Unknown word extraction for Chinese documents, in: Proc. The International Conference DBLP, 2002.

[6] G. Zou, Y. Liu, Q. Liu, Internet-oriented Chinese new words detection, Journal of Chinese Information Processing 18(6)(2004) 1-9.

[7] X. Yang, W. Yang, An analysis on the modern Chinese neologisms, Chinese Language Learning (2009).

[8] X. Huang, R.F. Li, Discovery method of new words in blog contents, Modern Electronics Technique 36(2)(2013) 144-146.

[9] M. He, C.C. Gong, H. Zhang, Method of new word identification based on lager-scale corpus, Compute Engineering & Applications 43(2-1)(2007)157-159.

[10] S.C. Shi et al. New word identification based on large-scale corpus, Journal of Shandong University 41(3)(2006) 43-45.

[11] Z. Lin, X. Jiang, A new method for Chinese new word identification based on inner-pattern of-word, Computer & Modernization 1(11)(2010) 162-164.

[12] S. Huo, M. Zhang, Y.-Q. Liu, S.-P. Ma, New words discovery in Microblog content, Pattern Recognition & Artificial

Intelligence 27(2)(2014) 141-145.

[13] H. Zhang, C. Peng, J. Luan, Rapid algorithm for left(right) entropy of character strings based on external sort, Computer Engineering & Applications 47(19)(2011) 18-20.

[14] A. Mccallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proc. Conference on Natural Language Learning at Hlt-Naacl Association for Computational Linguistics, 2003.

[15] P. Bheganan, R. Nayak, Y. Xu, Thai word segmentation with hidden Markov model and decision tree, in: Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining Springer Berlin Heidelberg, 2009.

[16] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, London, England, 1999.

[17] ICTCLAS. <http://ictclas.nlpir.org/>, 2016.

[18] C.D. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, London, England, 1999.