ICIC Express Letters
Part B: Applications
Volume **7**, Number **1**, January **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **167–172**

RESEARCH ON CLASSIFICATION OF TOPIC SENTENCES COMBINED WITH TERM RECOGNITION

Hui Liu[1] and Yao Liu[1,2]

[1]Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038, P. R. China
liuy@istic.ac.cn

[2]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research
No. 35, North Fourth Ring Road, Beijing 100101, P. R. China

Received June 2015; accepted August 2015

Abstract. *The abstract of patent literature contains topic sentences of type, domain, function, etc. The accurate classification of topic sentences benefits analysis of patent literature, construction of patent knowledge base and technology-effect matrix. Corpus is labeled manually with corresponding rules written by analyzing the characteristics of patent terms. Conditional Random Fields is used to train and test labeled data. Secondly, terms in abstracts are replaced by the same word and Support Vector Machine is used to classify topic sentences. Experimental results show that the precision of classification based on SVM and CRFs is 94.9% which is superior to methods based on other models.*
Keywords: Conditional Random Fields, Term recognition, Support Vector Machine, Classification

1. **Introduction.** The development of science and technology is reflected by patent documents. The speed of development can be increased by using patent documents effectively for a country. Getting important technology information out of patent documents is the prerequisite for using patent documents. The abstract of patent document can be seen as semi-structured text. An abstract can be divided into three parts according to the theme. The type and domain of a patent are introduced in the first part. The component of equipment and details of technique are introduced in the second part. The function and utility are introduced in the third part. Knowledge base can be built by using information extraction technology from the previous two parts [1]. Patent effect matrix for patent analysis can be formed by extracting information from the third part. Therefore, classification work should be finished before the information extraction. Due to the fact that there are no strict restrictions on the format, the abstract may lack the component part or utility part. Therefore, it is necessary to use classifier to divide abstracts into different categories.

Most text categorization works classify different texts into different categories. News websites, for example, set different categories according to the property of news texts [2]. Common text classification algorithms include Native Bayes classifier, maximum entropy classifier, etc. Literature [3] proposed a high performance method which employs a two-step strategy to classify texts. In the first step, authors regard the words with parts of speech verb, noun, adjective and adverb as candidate features, perform feature selection on them in terms of the improved mutual information formula, and classify the input texts with a Native Bayes classifier. In the second step, authors regard the bigrams of words with parts of speech verb and noun as candidate and use the same feature selection to deal with the texts in the fuzzy area. Literature [4] proposed an improved maximum entropy text classification, which fully combines c-mean and maximum entropy algorithm

advantages. The algorithm takes Shannon entropy as maximum entropy model of the objective function, simplifies classifier expression form and uses c-mean algorithm to classify the optimal feature. Literature [5] described how to recognize comparative sentences from text documents by applying rule-based methods and statistical methods. Literature [6] proposed a novel method which is called associative rule-based classifier aggregating with category similarity. The method adopts the modified chi-square statistical technique to extract feature terms from each category.

Unlike text categorization works mentioned before, the classification of a patent abstract classifies different sentences in the same abstract. In the short text classification, texts are hard to model and feature vectors are sparse [7]. On the other hand, different sentences in the same abstract describe the same thing and contain same patent terms. This paper proposes a method to classify sentences based on Conditional Random Fields used to recognize patent terms and support vector machines used to classify sentences after replacing all terms with same word.

2. **Term Recognition Based on Conditional Random Fields.** Conditional Random Fields which is proposed in 2001 by Lafferty is a conditional arbitrary undirected graphical model to label sequential data [8,9]. The contingent probability of observing sequence is used to predict the most possible labeled sequences. Conditional Random Fields brings together the best of generative models and Maximum Entropy Markov Models. Unlike linear-chain models, Conditional Random Fields can capture long distance dependencies between labels. For example, if the same name is mentioned more than once in a document, all mentions probably have the same label, and it is useful to extract them all, because each mention may contain different complementary information about the underlying entity.

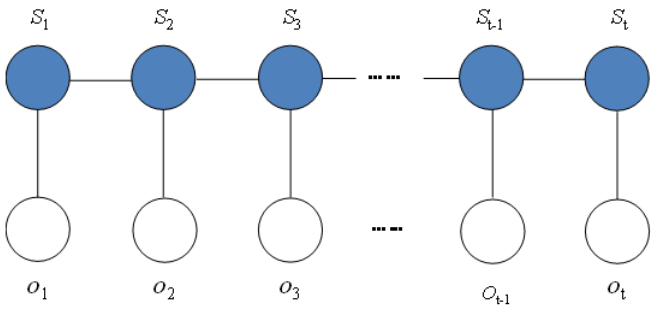

FIGURE 1. The structure of Conditional Random Fields

2.1. **The characteristic of patent term.** The term is the basic linguistic unit of scientific concept. A term can be a noun or a noun phrase, but it does not contain any stop words or pronouns. There is not a standard definition to distinguish terms from general words. The labeling method in this paper refers to *The Modern Technological Terms Dictionary*. The length of patent terms can be one word like "signal" to several words like "image signal processing device". Some common words like "method" or "cost" are not considered as terms. Patent terms have some characteristics. For example, there are many long terms like "frequency value progressive increase integrated circuit" and some abbreviations like "CDMA" or "ISDN". Long terms contain more ambiguities so that they are ambiguous and hard to label correctly by algorithms.

2.2. **Corpus processing of term recognition.** Sequence labeling method which uses signs "BIEO" was adopted. "B" was used to label the first word of a term, "I" was used to label the middle word, "E" was used to label the last word and "O" was used to label other words and punctuations in the abstract. The process of term recognition was regarded as

a labeling process. Words and their part of speech were regarded as feature. The Stanford part-of-speech tagger was used in the process. If the sentence is "The invention relates to a power supply circuit of an intelligent release", the result will be "The DT invention NN relates VBZ to TO a DT power NN supply NN circuit NN of IN an DT intelligent JJ release NN". Then "BIEO" was used to label terms in abstracts based on the corpus which was labeled by researchers. So the sentence will be "The DT /O invention NN /O relates VBZ /O to TO /O a DT /O power NN /B supply NN /I circuit NN /E of IN /O an DT /O intelligent JJ /B release NN /E". Every word was put on a single line with symbols like "The DT /O" and a blank line is entered between two abstracts.

3. **Topic Sentence Classification Based on Support Vector Machine.** Support Vector Machine is a machine learning technique developed on the basis of statistical learning theory, and it is one of the most successful realizations of statistical learning theory [10]. Support Vector Machine constructs an optimal hyper plane utilizing a small set of vectors near boundary and analyzes the consistency of learning and speed of convergence from structure risk minimization principle [11,12].

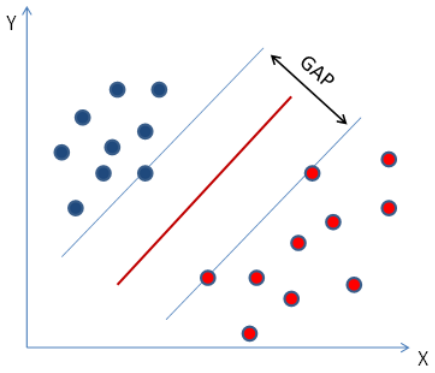

FIGURE 2. Example of Support Vector Machine

3.1. **The characteristic of topic sentences.** Type topic sentences are usually the first sentence of abstracts. Component topic sentences consist of hundreds of words and utility topic sentences consist of dozens of words. Different topic sentences in the same abstract describe the same thing. For example, the component topic sentences "The microprocessor includes a substrate bias rail providing a bias voltage during a first operating mode, a supply node providing a core voltage, a clamp device coupled between the bias rail and the supply node, and control logic" and the utility topic sentences "The control logic separately turns on and off clamp devices to selectively clamp the substrate bias rails in the first and second areas based on various power modes" contain same words like "substrate bias" and "clamp devices". Instead of being features, terms decrease the accuracy of classification. Therefore, all terms are replaced by the same word so that the effect on classifier can be reduced. On the other hand, many abstracts only include two types of topic sentences and topic sentences are not in a same order. Therefore, it is necessary to use classifier to identify the type of sentences.

3.2. **Corpus processing of topic sentence classification.** Firstly, terms recognized by Conditional Random Fields were replaced by the same word. For example, "The electronic system comprises a system carrier, a chip and an antenna, wherein the chip and the antenna are integrated on the system carrier, the antenna is an annular micro strip antenna, and the radio frequency signal receiving and transmitting ends of the chip are connected with two micro strip lines of an antenna feed-in signal" was revised as "The

term comprises a term, a term and an term, wherein the term and the term are integrated on the term, the term is an term, and the term receiving and transmitting ends of the term are connected with two term of an term". Secondly, stop words which are meaningless in the abstract were deleted. Every one of rest words can be seen as a feature in the classification. However, too many features will increase the calculation time and affect the accuracy of classification. So importance of words was calculated by using TF-IDF algorithm [13]. TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Top 50 words selected were seen as features and values of TF-IDF were seen as eigenvalues. Every topic sentence converts into a vector which uses first figure to indicate the category. If some top 50 words are not in a sentence, values of them will be zero. So a vector will be the form like "1 1:0.708 7:1056 14:−0.3333 19:0.1246 28:0.7823 33:0.1027 45:−0.2139".

4. **Experiment.**

4.1. **Data of experiment.** Two thousand abstracts in the field of communication were used in the experiment. Terms and categories were labeled by three researchers and the final result gets intersection elements of three label results. Two thousand abstracts contain 4216 terms and every abstract contains two or three categories of topic sentences. In the term recognition step, abstracts were labeled with "BIEO" and divided into five subsets to crosscheck.

```
Unigram
U000:%x[-2,0]
U001:%x[-1,0]
U002:%x[0,0]
U003:%x[1,0]
U004:%x[2,0]
```

FIGURE 3. A part of CRFPP template file

In the classification step, all terms were replaced by the same word "term". There are 1815 component topic sentences and 1604 utility topic sentences in abstracts. Topic sentences were put in different files according to their categories and feature vectors were calculated. 5-fold cross-validation method was also used in the classification step.

4.2. **Results of experiments and analysis.** CRFPP toolkit was used as the experiment tool for the term recognition. The important parameter C which represents the fitting degree was set to 2 after applying 5-fold cross-validation. The CRFPP toolkit uses train corpus to construct model and uses model to label test corpus with "BIEO". When words of a term are all labeled correctly by the toolkit, the term is labeled correctly.

Results show that Conditional Random Fields recognizes terms with a precision of 85.9 percent. Most complicated terms were recognized correctly like "downlink digital medium-frequency signal board". However, there are still some mistakes after comparing with results which are labeled by researchers. Mistakes focus in the following respects.

(1) Terms recognized are incompletion. For example, "outdoor air channel" was recognized as "air channel".

(2) Terms recognized contain redundant words. For example, releasing "driving circuit" was recognized as "releasing driving circuit".

Table 1. Results of term recognition

Number	P	R	F
A	0.852	0.813	0.832
B	0.868	0.824	0.845
C	0.886	0.802	0.842
D	0.833	0.789	0.81
E	0.857	0.802	0.829
Average	0.859	0.806	0.832

(3) Common words are recognized as terms. For example, "implement method" was recognized as a term.

(4) Complicated terms are recognized incorrectly. For example, "TCP transport protocols" was recognized as "TCP" and "transport protocols".

Hidden Markov Model and Maximum Entropy Model are used to compare with Conditional Random Fields. Same data are tested in these experiments.

Table 2. Results of contrastive experiments

Algorithm	P	R	F
Hidden Markov Model	0.809	0.756	0.782
Maximum Entropy Model	0.713	0.641	0.675

Results show that Hidden Markov Model recognizes terms with a precision of 8.9 percent, which is better than Maximum Entropy Model. However, Conditional Random Fields are more effective than these experiments with both higher accuracy and recall rates. Relative to other algorithms, Conditional Random Fields need more parameters and time to train in experiments. It has no need of strict independence assumption and can accommodate more information of context.

In the classification step, LIBSVM was used to classify topic sentences. Parameter C was set to 8 while Parameter G was set to 0.03125. LIBSVM uses train corpus to construct model and uses model to classify test corpus.

Table 3. Results of classification

Model	Precision
SVM+CRF	0.949
SVM	0.826
Naïve Bayes	0.737
K-NN	0.686

Results show that the classification of topic sentences combined with term recognition with a precision of 94.9 percent is 12 percent higher than the one without term recognition. Therefore, terms influence the precision of classification. Results also show that the effect of Support Vector Machine algorithm is better than others in the experiment.

5. **Conclusions.** The classification of topic sentences is achieved by using Support Vector Machine combined with Conditional Random Fields. Term recognition is accurate about 85.9 percent and Classification is accurate about 95 percent. Results show that the classification of topic sentences combined with term recognition can be very effective.

In real applications, simple algorithms are needed to locate the boundary of sentences, and then the method in this paper can be used to complete the task. In addition, replacing terms with the same word benefits the result of classification, but to know how the term recognition influences classification, there is still a lot of further work to do.

Acknowledgment. This work is partially supported by the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDD201502), National Key Project of Scientific and Technical Supporting Programs No. 2013BAH21B 02. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Liu, Z. Sui, Q. Zhao, Y. Hu and R. Wang, On automatic construction of medical ontology concept's description architecture, *International Journal of Innovative Computing, Information and Control*, vol.8, no.5(B), pp.3601-3616, 2012.

[2] X. Hua, F. Xu and Z. Wang, Fine-grained classification method for abstract sentence of scientific paper, *Computer Engineering*, vol.38, no.14, pp.138-140, 2012.

[3] X. Fan and M. Sun, A high performance two-class Chinese text categorization method, *Chinese Journal of Computers*, vol.29, no.1, pp.124-131, 2006.

[4] X. Li, Research of text categorization based on improved maximum entropy algorithm, *Computer Science*, vol.39, no.6, pp.210-212, 2012.

[5] C. Zhang, C. Feng and Q. Liu, Chinese comparative sentence identification based on multi-feature fusion, *Journal of Chinese Information Processing*, vol.27, no.6, pp.110-116, 2013.

[6] T. Feng, X. Gui and P. Yan, Associative rule-based text categorization method using category similarity, *Journal of Xi'an Jiaotong University*, vol.46, no.12, pp.6-11, 2012.

[7] J. Liu, Study on Chinese short message text classification based on theme, *Computer Engineering*, vol.36, no.4, pp.30-32, 2010.

[8] J. Lafferty, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, *Proc. of International Conference on Machine Learning*, San Francisco, CA, pp.282-289, 2002.

[9] F. Peng and A. McCallum, Information extraction from research papers using conditional random fields, *Information Processing & Management*, vol.42, no.4, pp.963-979, 2006.

[10] V. Vapnik, An overview of statistical learning theory, *IEEE Trans. Neural Networks*, vol.10, no.5, pp.988-999, 1999.

[11] T. Liu, Y. Yang, H. Wan, Q. Zhou and B. Gao, An experimental study on large-scale web categorization, *Posters Proc. of International World Wide Web Conference*, pp.1106-1107, 2005.

[12] J. Su, B. Zhang and X. Xu, Advances in machine learning based text categorization, *Journal of Software*, vol.17, no.9, pp.1848-1859, 2006.

[13] Y. Yang and J. Pedersen, A comparative study on feature selection in text categorization, *ICML*, vol.97, pp.412-420, 1997.