Research on the Performance of Segmentation of Text Classification Based on CNICC

Jun-peng Gong
School of Computer of Science
Communication University of China
Beijing, China

Jia Song
School of Computer of Science
Communication University of China
Beijing, China

***Abstract*—With the development of Internet, people have entered the era of New Media, and the demand for intelligent information is more intense. How to make news industry, provide more personalized and efficient service for all walks of life, is something worth exploring. In this paper, we aim to analyze the performance of segmentation, which is the foundation of text classification. Our study describes the method of creating a type thesaurus about news based on Chinese news information classification and code. Experimental results show that the approach is a valid and promoted method of segmentation optimization in news classification.**

Keywords—text mining; Chinese news information classification and code; news classification; type thesaurus;segmentation

I. INTRODUCTION

In recent years, by the development and wide popularity of media information technology, more and more news organization has changed their mode of production. Among these, text automatic classification[1][2] is very popular. Hence, how to efficiently develop a high efficient system for news classification from unorganized massive news data is becoming a hot research topic. This paper aims to employ the idea of segmentation optimization[3][4] in news classification, and apply the method of constructing a thesaurus, which have sourced from different vocabulary databases, to improve the performance of segmentation in the process. By experiments, our results show that constructing thesaurus can inhance the rationality of text segmentation in news classification.

The rest of this paper is organized as follows. In section 2, some related work (in the content of Chinese News Information Classification and Code aspect) relevant to our study is presented. Section 3 introduces the method of constructing a type thesaurus and the approach of using the thesaurus in the process of text segmentation. Experiments, results and some analysis are shown in section 4. Finally, our conclusions and direction for future work are given at the end of this paper.

II. RELATED WORK

A. The brief description of CNICC

CNICC, short term for Chinese News Information Classification and Code[5]-[8], is a national standard for china news industry. It is useful for news information exchanging, storing, processing and sharing between news industries and users. The standard, which reflects the essence of news, absorbs the advantage of domestic and foreign news industry methods about news classification, such as IPTC, Xinhua News Agency, People's Daily, and so on. At present, most of china news organizations use the standard for news classification.

B. The Content of CNICC

CNICC is composed of basic class (also known as the first class directory), simple table, detailed table, complex sub table (also called auxiliary table) and instructions. According to idea of news theme, the basic class divid news into 24 categories, which are political, legal(or justice), foreign relations, military, social work, accident(or disaster), economy, finance, infrastructure(or real estate), agriculture(or rural areas), mining(or industry), energy(or water conservancy), electronic information industry, transportation(or postal services), commerce(or foreign trade and customs), service industry(or tourism), environment(or meteorology), education, science and technology, culture(or entertainment), literature(or art), media, medicine(or public health) and sports. Based on the basic large class, the stepwise subdivision is carried out, and the subdivision level can reach five levels. Besides, there are 331 categories in the second layer, and the total number of more than 4000 categories in all levels. Each class is expressed by the code, the category name and description. These concrete contents can refer to GB/T 20093-2006, called Chinese News Information Classification and Code.

III. THE APPROACH OF CONSTRUCTING THESAURUS

Due to news have a large number of professional vocabularies, most segmentation tools own words which based on the mass leading to the professional level is not high. As a result, some defects still exist in the specific field of text classification (such as news) on the performance of the segmentation. Thus, we should establish a thesaurus of news cagtegories in order to improve the word segmentation process. The following are specific methods.

A. Chinese Information Classification and Codes

Considering the standard has a wide coverage, in which the category name and the instruction contains many thesaurus categories information, so we can regard the category name and the instruction as a part of the thesaurus. In the process of

ICIS 2016, June 26-29, 2016, Okayama, Japan

selecting words, it should pay attention to three points. First, the category name contains a group of words, it should divide into separate word, such as foreign institutions and foreign affairs divided into two lexicon. Second, the keyword about description of the category belongs to the thesaurus. Last, Theme words, not duplicated in the category system, belongs to the thesaurus.

B. Sogou Vocabulary Database

Sogou vocabulary database[9], based on its search technology and input method, is a series of professional thesaurus. It contains information about the city, natural science, social science, engineering, agriculture, forestry, medicine, electronic games, art and design, sports and leisure, humanities, entertainment, and so on. Therefore, selecting some terms about news as a part of the thesaurus is feasible. In the process of selecting words, it should pay attention to one aspect, which is that the corresponding words of the low classification precision and recall should be selected as much as possible.

C. A classified Dictionary of Media English

The media english dictionary[10] is an early tool book, which is used for english language worker listening to radio, reading english newspapers and magazines. It collected nearly million news common words about political, military, law, economy, culture, society, life, education, sports, science and technology. Using terms in it as a part of the thesaurus is undoubtedly representative. In the process of selecting words, it should pay attention to two points. The first is not to copy the chinese translation of words, but to select the term, has on behalf of the category. The second is the selected word should be in accordance with the rules of the standard, duing to the english dictionary has only eight categories and the standard has twenty-four categories.

IV. EXPERIMENTAL RESULTS

In order to illustrate the effectiveness of our approach, which adopts the method of constructing the thesaurus, we construct the below experiment. The flow chart of news classification is shown Fig.1.

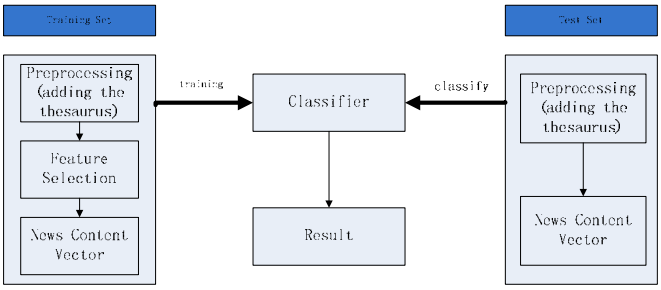

Fig. 1. The flow chart of news classification

The experimental results shown in Table 1-Table 4.

TABLE I. THE PERFORMANCE OF PRECISION

Type	precision	
	Not adding the thesaurus	***Adding the thesaurus***
medicine	87.3	87.8
mining industry	88.1	90.1
disaster	96.1	96.1
law	71.9	73.5
service industry	78.5	83.1
science and technology	84.4	91.1
real estate	72.6	79.6
energy	78.9	77.3
social	68.8	68.2
transportation	80.2	81.4
culture	68.9	66.7
environment	78.3	83.8
politics	60.9	64.4
electronic	78.2	84.2
international relations	51	53.2
education	81.3	85.2
literature and art	73	75
economics	62.4	63
commercial	72.7	75.6
sports	94.2	95
agriculture	71.7	70.3
media industry	83.1	83.5
military	76.8	76.9
finance	81.8	80.4

TABLE II. THE PERFORMANCE OF RECALL

Type	recall	
	Not adding the thesaurus	***Adding the thesaurus***
medicine	83	86
mining industry	89	91
disaster	99	99
law	82	83
service industry	73	69
science and technology	54	51
real estate	77	78
energy	86	85
social	53	60
transportation	89	83
culture	71	72
environment	83	83
politics	56	58
electronic	86	85
international relations	73	74
education	91	92
literature and art	84	84
economics	53	63
commercial	56	62
sports	98	96
agriculture	66	71
media industry	69	76
military	73	80
finance	90	90

TABLE III. THE PERFORMANCE OF F1

Type	F1	
	Not adding the thesaurus	***Adding the thesaurus***
medicine	85.1	86.9
mining industry	88.6	90.5
disaster	97.5	97.5
law	76.6	77.9
service industry	75.6	75.4

science and technology	65.9	65.4
real estate	74.8	78.8
energy	82.3	81
social	59.9	63.8
transportation	84.4	82.2
culture	69.9	69.2
environment	80.6	83.4
politics	58.3	61.1
electronic	81.9	84.6
international relations	85.8	88.5
education	60.1	61.9
literature and art	78.1	79.2
economics	63.3	68.1
commercial	57.3	63
sports	96.1	95.5
agriculture	68.8	70.6
media industry	75.4	79.6
military	74.9	78.4
finance	85.7	84.9

TABLE IV. THE COMPREHENSIVE PERFORMANCE OF NEWS CLASSIFICATION

Index	Not adding the thesaurus	Adding the thesaurus
macrorecall	76.4	78
macroprecision	76.7	78.6
macrof1score	76.1	77.8
accuracy	76.4	78

In the experiments, reducing the influence of the data imbalance, we specify that texts are equally distributed across different categories, for example, specifying 1500 messages for each of the 24 categories. We use another 100 messages as the testing data.

V. CONCLUSIONS

In this paper, we introduce the application of constructing the thesaurus in news classification. Through the experiments, we conclude that the effectiveness of applying the method in the automatic information processing of news. Besides, we, on the whole, discover that performance indexes such as precision, recall, F1, macroprecision, macrorecall, macrof1score and accuracy have improved. In the future, we would like to improve our approach, in particular filtering the vocabularies of similar categories will be considered in news classification.

ACKNOWLEDGMENT

This research was financially supported by the National Key Technology R&D Program 2014BAK10B01.

REFERENCES

[1] WEI Xiaoning, ZHU Qiaoming and LIANG Xingyan, in: Using Bayesian in Text Classification with Par ticiple- method. Journal of Suzhou Vocational University Vol.19 No.1 (2008).

[2] Cui Caixia and Zhang Zhaoxia, in: Contrast Research on Text Categorization Methods. Journal of Taiyuan Normal University(Natural Science Edition) Vol.6 No.4 (2007).

[3] Zhang R Q,Yasuda K,Sumita E.Chinese Word Segmentation andStatistical Machine Translation. ACM Transactions on Speechand Language Processing(TSLP) . 2008.

[4] Nianwen Xue.Chinese Word Segmentation as Character Tagging. Computational Linguistics . 2003.

[5] Wang Wenwen: The Application of Chinese News Information Classification and Codes in the News Reference Room.Library No.5(2007).

[6] Xu Man: Study on Standard of Chinese News Information Classifying[D]. Wuhan University Press(2005).

[7] Zhang Zhi-ping: Text Classification Based on Chinese News Information Classification and Code. Journal of Taiyuan University of technology Vol.41 No.4(2010).

[8] Deng qian, Ling hong: Conception and implementation of automatic indexing and classification for chinese news information. Science technology for china's mass media Vol.9 No.115 (2005).

[9] http://pinyin.sogou.com/dict/.

[10] Lin Mei: A classified Dictionary of Media English. Book. Foreign Language Teaching and Research Press.