New Word Identification for Chinese Patents Based on Multiple Statistic Measures and Pattern Combination

Xiong Wen[1, 2]

[1] Beijing Normal University, Institute of Chinese Information Processing, Beijing 100875, China

[2] The China Patent Information Center, the State Intellectual Property Office, Beijing 100088, China

Email: stevens7979@sina.com

Keywords: New Word Identification (NWI); Out of Vocabulary (OOV); Pattern Combination; Candidate Generation; Statistical Measures Integration; Pattern Filtering

Abstract. New Words Identification (NWI) is one of the critical researches in Chinese Natural Language Processing (NLP), which has important influence to the successive tasks of Chinese NLP. Aiming at the problem of the NWI, which is disturbed in the automatic or half-automatic processing for text translation of Chinese patents, this paper proposed a method for NWI of Chinese patents based on integration of multiple statistic measures and pattern combination, which included a specifically preprocessing method for string dividing, where the technological terms in patents were reserved, and non-technological words were removed as many as possible; then, the divided strings with different lengths were combined using multiple patterns with a greedy maximum match to generate candidates; furthermore, the noisy candidate strings were filtered using four filtering patterns summarized manually; finally, the statistical measures only adapting to two variables were extended to those adapting to multiple ones; in the meantime, the values of the multiple statistic measures extended were integrated by using a ranking method, which evaluated the candidates according to the thresholds to form the set of new words. Experiments on abstract texts of Chinese patents showed that the precision can reach 80%; and the F1 value can reach 68.15%, verifying the effectiveness of the method.

Introduction

Along with the bringing forward of the innovative ability of the nation, the number of patents applying is fast increasing in China now. Therefore, automatically or half automatically handling the huge Chinese patents becomes a rising requirement. Chinese segmentation is a key and basic step of the preprocessing of the patent texts. New words in patents are kind of words Out of Vocabulary (OOV), which are needed to be automatically or half automatically recognized and extracted. Therefore, New Word Identification (NWI) is an unavoidable process in Chinese segmentation.

The method using vocabulary for Chinese segmentation is used for most occasions due to its fast, controllable, and convenient. In this method, OOVs usually form fragments of segmentation, which influence the following processes of the patent texts, including Machine Aided Human Translation (MAHT) of Chinese patents, Machine Translation (MT), keywords automatically indexing, the whole text retrieve. Therefore, NWI in patent texts is significant for these automatic processes.

Many NWI methods are appeared in previous literature. However, these methods cannot be directly transplanted to the patent field due to the characteristics such as the strong technology, low frequency appearing, and random distribution. Meantime, many especial expression models of the sentences are used in patent texts, which increase the difficulty of NWI in patents. Aimed at this difficulty, this paper presents a method for NWI in patent texts based on integration of the multiple statistic measures and pattern combination, which is effective.

The rest of the paper is organized as follows: first, in Section 2, the new method for NWI in patent texts is presented; then, in Section 3, the experiments were described, where the experimental data and results were exhibited; and then, in Section 4, the related work for the NWI is introduced in brief; finally, in Section 5, conclusions are reached from the results of the experiments.

The New Method for NWI in Patents

The patents applied can be divided into three categories, i.e. invention, utility model, and appearance design, where the abstract texts of invention include more contents of high technology than the others, which is a main data source handled by the paper. The new method for the NWI in patents was composed by five steps as follows.

Segmentation Step: Segmentation for the abstract texts of the invention patents using a term lexicon in patent field can distinguish more technological terms than using a lexicon for general purposes, where lots of terms are daily words with little technological info. Therefore, the selection of segmentation lexicon is an important precondition, which will be described in the Section 3.

Process for Chinese scattering string Step: To reduce the noise in the recognition procedure, we need a special preprocessing algorithm, which got rid of the non-tech words (which are noisy fragments when forming new word candidates in patents.) Meantime, to identify new words in patents as many as possible, we need to reserve tech words and tech fragments as many as possible (which are useful fragments to form new word candidates in patents.) Therefore, we need to generate a general lexicon used for filtering the non-tech words. Considering the big difference in language custom between the People Daily corpus and patent texts, we extracted high frequency words in People Daily corpus to compose the general lexicon for the filtering purposes. On the other hand, the algorithm segmented the Chinese sentences into the sequences of scattering strings using lexicon of stop word and explicit tokens such as punctuations, English numbers, English chars, and other symbols which are not Chinese chars.

Candidate generation for NWI in patents Step: The adjacent scattering strings in Chinese formed the material of NWI in patents due to that non-tech info was filtered in step 2. These scattering strings were divided into four categories according to their lengths such as words with a single Chinese char, words with two Chinese chars, words with three Chinese chars, and words with four or above Chinese chars. Therefore, new words can be formed by combining the adjacent scattering strings according to different combination patterns as Tab. 1.

Table 1. Four patterns to form candidates

Patterns	Instances	Num.
Single Chinese Char	c_1c_1 c_1c_m $c_1c_1c_m$ c_2c_1 $c_2c_1c_1$ $c_2c_1c_m$ c_3c_1 $c_3c_1c_m$	8
Two Chinese Chars	c_2c_m	1
Three Chinese Chars	c_3c_m	1
Four or above Chinese Chars	c_4c_m	1

Where: c_1, c_2, c_3, c_4, and c_m stood for a Chinese word which is formed by a Single Chinese char, two Chinese chars, three Chinese chars, and four or above Chinese chars correspondingly. Using these elements, the candidates can be composed by them according to the instances of the patterns with a maximum match from left to right.

Special patterns for filtering Step: By analyzing the candidate words, we found that the candidates cannot be new words in patents if some special patterns appeared in these candidates. Therefore, we concluded four patterns for new word filtering such as patterns starting with special char or words, middle patterns, ending patterns, and starting and ending patterns. The samples of these four patterns are illustrated as Tab. 2.

Table 2. The samples of the four filtering patterns

Filtering Patterns	Instances	Num.
Starting	相(xiang), 设(she), 化(hua), 端(duan), 无(wu) 纯(cun), 有(you), 内(nei), 适(shi), 呈(cheng)	65
Middle	和(he), 后(hou), 一(yi), 内(nei)	4
Ending	设(she), 煎(jian), 呈(cheng), 大(da), 达(da) 低(di), 加(jia), 相(xiang), 身(shen), 侧(ce)	42
Starting and Ending	伸(shen), 设(she), 受(shou), 置(zhi), 接(jie) 扣(kou), 通(tong), 埋(mai), 装(zhuang), 控(kong)	59

Integration filtering with multiple statistic measures Step: To refine the results from the step 4, we can use statistic measures as filtering thresholds to get rid of the noisy candidates. As known, single statistic measure can pop out a certain statistic law. For instance, Mutual Info (MI) measure can effectively handle the concurrence of the high frequency, but it cannot when the frequency is low. On the other hand, the Log likelihood measure can be useful in the occasion of low word frequency. Therefore, the vote method was adopted, which evaluated the candidates of new words using the integration of multiple statistic measures, and then, a threshold was set to filter the candidates with low scores.

As known, the MI, Dice, and Chi-square are used for evaluations on the two combinative parts formed the whole. However, when the number of the combinative parts is greater than two such as $c_1c_1c_m$, and $c_2c_1c_m$, we need to extend these original statistic measures to adapt them to the multiple parts. The MI measure was extended according to the Eq. 1.

$$M(w_1...n) = \log(\frac{p(w_1...n)}{Avp}) \quad s.t. \quad Avp = \frac{1}{n-1} \cdot \sum_{i=1}^{n-1} p(w_1...i) \cdot p(w_{i+1}...n). \tag{1}$$

Where: $p(w_1...w_n)$ is the word frequency formed by multiple Chinese chars. And we extended the Log likelihood as the Eq. 2.

$$L(w_{1...n}) = \frac{1}{n-1}\sum_{i=1}^{n-1} L_i(w_{1...i}, w_{i+1...n})$$
$$s.t.\ L_i(w_{1...i}, w_{i+1...n}) = 2 \cdot (a_i \cdot \log_2(\frac{a_i \cdot N}{(a_i+b_i)\cdot(a_i+c_i)}) + b_i \cdot \log_2(\frac{b_i \cdot N}{(a_i+b_i)\cdot(b_i+d_i)}) + c_i \cdot \log_2(\frac{c_i \cdot N}{(c_i+d_i)\cdot(a_i+c_i)}) + d_i \cdot \log_2(\frac{d_i \cdot N}{(c_i+d_i)\cdot(b_i+d_i)})) \tag{2}$$

Where: a word formed by multiple Chinese chars can be split to two differently combinative parts, the whole combinations of which can be calculated according to the original Log likelihood, and an average can be taken as the final value wherein the four factors such as a_i, b_i, c_i, and d_i were defined as Tab. 3.

Table 3. Four Factors in Eq. 2

Appearing	Back word	No back word
Front word	a_i	b_i
No front word	c_i	d_i

And the *N* in Eq. 2 is equal to $a_i + b_i + c_i + d_i$. Similarly, we extended the Chi-square as the Eq. 3.

$$\chi^2(w_{1...n}) = \frac{1}{n-1}\sum_{i=1}^{n-1} \chi^2{}_i(w_{1...i}, w_{i+1...n})$$
$$s.t.\ \chi^2{}_i(w_{1...i}, w_{i+1...n}) = \frac{a_i \cdot d_i - b_i \cdot c_i}{(a_i+b_i)\cdot(a_i+c_i)\cdot(b_i+d_i)\cdot(c_i+d_i)} \tag{3}$$

And we extended the Dice coefficient as the Eq. 4.

$$D(w_{1...n}) = \frac{1}{n-1}\sum_{i=1}^{n-1} D_i(w_{1...i}, w_{i+1...n})$$
$$s.t.\ D_i(w_{1...i}, w_{i+1...n}) = \frac{2 \cdot a_i}{(a_i + b_i) \cdot (a_i + c_i)} \quad (4)$$

We also used the Term Frequency and Inverse Document Frequency (TF-IDF) for the evaluation of the candidates, which was calculated according to the Eq. 5.

$$Tfidf = tf(a) \cdot \log(N / df(a)) \quad (5)$$

Where: *tf*(a) is the frequency of the candidates, and the *df*(a) is the document frequency of them. The *N* is the total number of the documents.

We presented an integrative rank method, which first used the five statistic measures to rank the candidates, and then recorded the rank orders, finally calculated the integrative values according Eq. 6, where the rank of Log likelihood was sorted ascending, and the others were sorted descending, indicating the possibility that a candidate becomes a new word.

$$R = \sum_{i=M,L,\chi^2,D,Tfidf} \frac{1}{R_i} \quad (6)$$

Where: R_i is the rank order formed by the *i*th statistic measure. We sorted the candidates descending using the *R* value again, and filtered the candidates with the low scores.

Experiments

The forward maximum segmentation was adopted due to its fast and controllable, and the lexicon in this algorithm was formed by China Patent Info Center (CPIC) having 1,790,000 terms. The stop-word lexicon included 700 items, and the general lexicon was generated from People Daily corpus included 13,314 items, which were of high frequency such as China, development, economy, working, and task. The process procedure of a sentence was illustrated as follows.

Sentence: 纺粘长丝优选被形成皮－芯型结构，苯并咪唑和对苯二酰二氯的共聚的共聚物。 (The spun bond filaments are preferably formed in a sheath-core configuration, copolymer derived from the copolymerization of Benz imidazole and terephthaloyl dichloride.)

After segmentation, the Chinese sentence was divided into fragments as follows.

纺 粘 长丝 优选 被 形成 皮 － 芯型 结构 , 苯并咪唑 和 对苯二酰 二氯 的 共聚 的 共聚物 。

Where: the words and tokens of three categories were filtered as follows.

(i) Stop words such as Bei (被), He (和), and De (的).

(ii) General words such as YouXuan (优选), XingCheng (形成), and JieGou (结构)

(iii) Explicit tokens such as hyphen, comma, and full stop.

After gotten rid of these words and tokens, the sequence of scattering string from the sentence was as follows.

Sequence: 纺 粘 长丝 … 皮 … 芯型 … 苯并咪唑 … 对苯二酰 二氯 … 共聚 … 共聚物

Where, the ellipsis indicated some words or tokens were filtered. Therefore, the candidates must not span the ellipsis. Combining the adjacent scattering strings with special patterns, the candidates were formed as follows.

$c_1c_1c_m$: 纺＋粘＋长丝 (spun bond filament)

c_4c_m: 对苯二酰＋二氯 (terephthaloyl dichloride)

After calculating the extended integration value of multiple statistic measures, we gave two new words according to the conditions as follows, where the threshold of *R* was 0.05.

Table 4. The Statistic values and R values of the two new words

Measures	氨磺酰基+团 (sulphonamide group)	紧线+轮 (thread tension disk)
Extended mutual info	9.49995	8.11365
Extended Log likelihood	0.00001	0.00001
Extended Chi-square	0.01010	0.00505
Extended Dice	2.00000	1.00000
TFIDF	0.00071	0.00213
R	2.00530	0.05008

In experiment one, we randomly extracted 100 Chinese abstracts of invention as experimental corpus from the big part *D* (weaving and paper making) of Patent Cooperation Treaty (PCT) from year 2006 to 2012, and totally 22,175 Chinese chars. The evaluative method used the popular measures such as precision, recall, and F1, which were listed as Eq. 7.

$$P = tp/(tp+fp)$$
$$R = tp/(tp+fn) \quad (7)$$
$$F1 = 2\cdot tp/(2\cdot tp+fn+fp).$$

Where, *tp* is the true positive instances if the positive instances are judged as positive categories; *fp* is the false positive instances if the negative instances are judged as positive categories; *fn* is false negative instances if the positive instances are judged as negative instances. To calculate the results, we tagged the 100 abstracts, and obtained 271 new words in patents. The results of automatic recognition were illustrated as Fig. 1.

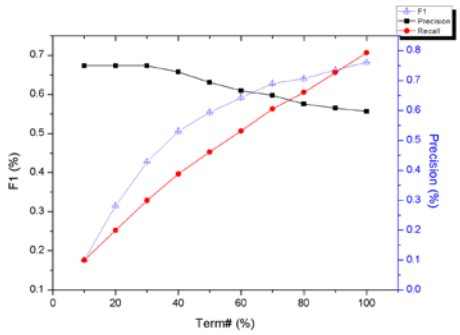

Figure 1. Precision, recall, and F1 of NWI from 100 abstracts

Where: the recall increased apparently and the precision decreased slightly when the number of the new words increased. The F1 value reached 68.15%, indicating that the method can work in a limited scale of corpus.

For validating the effectiveness of the algorithm when it had more data, experiment two randomly extracted 9,572 Chinese abstracts of invention as the experimental corpus from big part *D* of PCT from year 2006 to 2012, and totally 2,100,000 Chinese chars. We only evaluated the new words of top 300 of this experiment according to the precision due to the number of the candidates were large, and the result was listed as Tab. 5.

Table 5. The precision of top 300 words from the big part D with 9,752 abstracts

@number of candidate	P@100	P@200	P@300
Precision (%)	71.0	78.0	75.0

Where: the precision of the new words reached 78% in the experiment, indicating the precision of top 200 had an increase when the scale of the experimental corpus was increased.

For validating the effectiveness of the algorithm on other big parts, experiment three randomly extracted the abstracts of other seven big parts as the experimental corpus from year 2006 to 2012. We only adopted the precision to evaluate the new words of top 100 due to the candidates were large, and the result was listed as Table 6.

Table 6. The precision of top 100 words from other big parts

Big part no.	Abstract no.	Precision (%)
A	94651	73.0
B	89481	80.0
C	115149	69.0
E	11311	79.0
F	47630	79.0
G	107990	62.0
H	154159	73.0

Where: the highest precision reached 80.0%, indicating the precision of new words of top 100

increased when the scale of the corpus was increased. Some of the instances of new words according to the different pattern combination were listed as Tab. 7.

Table 7. The instances of new words using different combination patterns

No.	Combination patterns	New words
1	C_1C_m	聚＋对苯二甲酸 (terephthalate)
2	C_2C_m	筒状＋针织物 (tubular fabric)
3	C_2C_1	织物＋层 (layer of textile)
4	$C_2C_1C_1$	织造＋纤＋网 (woven web)
5	$C_2C_1C_m$	耐压＋缩＋疲劳性 (resistance to compression fatigue)
6	C_3C_m	不饱和＋单体 (unsaturated monomer)
7	C_3C_1	亚丙基＋酯 (polytrimethylene)
8	$C_3C_1C_m$	丙烯腈＋系＋聚合物 (acrylonitrile polymer)

Related work

Nearly 60 percent failure of the Chinese segmentation is caused by OOVs [1], which can be almost resolved by NWI, indicating the NWI is an important process of NLP in Chinese. According to the occasions, the methods of NWI can be divided into three categories: (i) instead of Chinese segmentation, n-grams language model was used for the NWI [2]; (ii) depending on the results of segmentation after using a certain linguistic tool, then, the NWI was executed. Hidden Markov Model (HMM) was used for tagging the semantic roles of the sentences, then, the new words were obtained according to the combinations of these tags [3]; (iii) combining the segmentation tool and the procedure of the NWI, the NWI was processed and the segmentation was handled at the same time [4].

On the other hand, according to the characteristics of the NWI, those methods can be divided into two categories [5]: (i) based on linguistic rules [6]; (ii) based on the machine learning and statistics. The concurrency frequency of the morphs of the new words was used for determining the border of the candidates, and then a filtering step was processed using syntactic and rule [7]. Up to now, hybrid methods were used for NWI, which integrated the rules and statistics to exert the superiority of the both [8, 9].

In [10], the syntactic parser was used for determining the possibility of the new words. Considering the significance and the difficulty of the NWI on the Micro-blog, the literature [11] put forward an iterative algorithm of context entropy using word association info, which obtained a list of candidates through the context relation, introducing the morphological characteristics and the word frequency for filtering to improve the F-value of the recognition.

Based on that there were lots of sentimental new words wrongly written (deliberately) on Micro-blog, and the style of the language was colloquial, non-restraint, and nonstandard, the paper [12] presented a recognition method for sentimental new words based on word vectors, which used the tool of open sources of Google named *word2vec* for the latent semantic relationship. In the tool, a neural network was used for the training of the word vectors, which got rid of the dependence on the external semantic resources for the sentimental computation.

Conclusion

The paper only used heuristic patterns for generation on the scattering strings, pattern filtering, and the extended statistic measures for rank, where the Part of Speech (POS) tagging, roles tagging, and syntactic parser was not used due to their heavy cost. The results of the method were

encouraging, and the method overcame the difficulties such as the training dependence on the corpus, the field relativity, the hugely computational cost, and the long computational time, which can be used as an online usage. In the future, the scope of the NWI will be extended to other parts of the Chinese patents such as the specification, claim, and technology background.

Acknowledgement

The paper was supported by “the National High Technology Research and Development Program of China (No. 2012AA011104)” and “the Fundamental Research Funds for the Center Universities.”

References

[1] Sproat, R. and Emerson, T. The first international Chinese word segmentation bakeoff [C]. The second {SIGHAN} workshop on Chinese language processing, Sapporo, Japan: 2003.

[2] Nie JY, Hannan ML, Jin WY. Unknown word detection and segmentation of Chinese using statistical and heuristic knowledge [J]. Communications of COLIPS, 1995, 5(1-2): 47~57.

[3] Zhang HP, Liu Q, Zhang H, Cheng XQ. Automatic recognition of Chinese unknown words based on roles tagging [C]. The 1st SIGHAN Workshop on Chinese Language Processing, Taipei, Taiwan-China: 2002.

[4] Peng FC, Feng FF, McCallum A. Chinese segmentation and new word detection using conditional random fields [C]. The 20th Int'l Conf. on Computational Linguistics (COLING 2004), Geneva, Switzerland: 2004.

[5] Zhang HJ, Shi SM, Zhu CY, Huang HY. Survey of Chinese new words identification [J]. Computer Science, 2010, 37(3): 6~10.

[6] Liang T, and Ye DY. Apply the word-building rules and the analogically neural network to the new words extraction [C]. Research on Computational Linguistics Conference XIII, Taipei, Taiwan-China: 2000.

[7] Chen K-J, Ma W. Unknown word extraction for Chinese documents [C]. COLING 2002, Taipei, Taiwan-China: 2002. (In Chinese)

[8] Wu A, Jiang Z. Statistically-enhanced new word identification in a rule-Based Chinese system [C]. The Second Chinese Language Processing Workshop, Hong Kong, China: 2000.

[9] Liu H. A new approach for domain new words detection [J]. Journal of Chinese Information Processing, 2006, 20(5): 17~23.

[10] Wu A. Chinese word segmentation in MSR-NLP [C]. The Second SIGHAN Workshop on Chinese Language, Sapporo, Japan: 2003.

[11] Huo S, Zhang M, Liu YQ, and Ma SP. New words discovery in micro-blog Content. PR & AI. 2014, 27(2): 141-145.

[12] Yang Y, Liu LF, Wei XH, and Lin HF. New methods for extracting emotional words based on distributed representations of words [J]. Journal of Shandong University (Natural Science), 2014, 49(11): 51~58.