SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2016, Vol. 59 092102:1–092102:11
doi: 10.1007/s11432-015-0906-9

A novel unsupervised method for new word extraction

Lili MEI, Heyan HUANG*, Xiaochi WEI & Xianling MAO

Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Department of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081*, China*

Received November 8, 2015; accepted January 6, 2016; published online August 11, 2016

Abstract New words could benefit many NLP tasks such as sentence chunking and sentiment analysis. However, automatic new word extraction is a challenging task because new words usually have no fixed language pattern, and even appear with the new meanings of existing words. To tackle these problems, this paper proposes a novel method to extract new words. It not only considers domain specificity, but also combines with multiple statistical language knowledge. First, we perform a filtering algorithm to obtain a candidate list of new words. Then, we employ the statistical language knowledge to extract the top ranked new words. Experimental results show that our proposed method is able to extract a large number of new words both in Chinese and English corpus, and notably outperforms the state-of-the-art methods. Moreover, we also demonstrate our method increases the accuracy of Chinese word segmentation by 10% on corpus containing new words.

Keywords new word extraction, word segmentation, domain specificity, statistical language knowledge, domain word extraction

Citation Mei L L, Huang H Y, Wei X C, et al. A novel unsupervised method for new word extraction. Sci China Inf Sci, 2016, 59(9): 092102, doi: 10.1007/s11432-015-0906-9

1 Introduction

Chunking is a very important basic precursor in both Chinese and English natural language processing (NLP) tasks. However, sentence chunking does not perform well on informal texts, e.g., Weibo and BBS, which is mainly caused by widely distributed new words [1].

In Web 2.0 based social media, new words are emerging in most languages and most domains everyday. Homophonic words, typos, abbreviations and domain-specific words and phrases are common phenomena in user-generated content. For example, in Chinese, some widely used homophonic words and abbreviations evolve into new meanings on social websites, and even the existing word may have different new explanation. Table 1 shows some typical instances, the Chinese new word "女票 (girlfriend)" is the abbreviation of "女朋友 (girlfriend)".

New word extraction is indispensable to many NLP tasks such as sentence chunking [2,3] and sentiment analysis [4]. However, automatic new word extraction is a challenging task. The reasons are as follows: (1) new words often have no fixed language pattern, appearing in a new form; (2) many new words appear

* Corresponding author (email: hhy63@bit.edu.cn)

Table 1 Examples of new words

New word	Origination	English translation
女票	女朋友	Girlfriend
涨姿势	长知识	Knowledge have been increased
高富帅	个子高、富有、帅气	A tall, rich and handsome man
走召弓虽	超强	Very strong

with the new meanings and usages of existing words; (3) it is very difficult to identify low-frequency new words and filter high-frequency garbage words.

Existing methods for new word extraction have made significant progress. However, these methods are suffering from poor flexibility and portability [5–7], or they can hardly capture special features of new words [8–10]. To address these shortcomings, we consider domain specificity and combine with multiple statistical language knowledge to extract new words. Our main ideas are as follows: (1) It is intuitive that new words in source domain (SD), e.g., new network words and academic words of computer science, rarely appear in a different domain named as reference domain (RD), e.g., the News corpus. Therefore, we extract n-grams from SD and perform a filtering algorithm based on domain specificity to obtain a candidate list of new words. (2) For a candidate new word, the statistical language knowledge can be used to quantify the possibility of being a new word. So we introduce a ranking method for new word extraction, considering word features like word frequency feature, word internal feature and neighborhood feature. Experimental results show that our method can effectively extract a large number of new words both in Chinese and English corpus.

The advantages of the proposed method are: a) it is fully unsupervised which avoids the time-consuming labeling procedure, b) it requires no linguistic resources, c) no manually defined rule is needed to filter out undesirable words, d) it is a language-independent and domain-independent method. The main contributions of this paper are summarized as follows:

- We propose a novel approach to the task of new word extraction.
- We demonstrate the effectiveness of statistical language knowledge, such as string frequency, string cohesion and string liberalization, in the task of new word extraction.
- Experiments show that our proposed method increases the accuracy of Chinese word segmentation by 10% on corpus containing new words.
- We demonstrate that our approach is language independent and domain independent; it can effectively extract new words from different domain-specific corpus in both Chinese and English.

The remainder of the paper is structured as follows: Section 2 summarizes the related work. In Section 3, we describe our method in detail. Then, the experimental results and discussions are presented in Section 4. Finally, we conclude our work and suggest future work.

2 Related work

Extensive work has been done on new word extraction, which can be categorized into rule-based methods, statistical methods and hybrid methods. Our work falls into the statistical category.

In the rule-based methods, new words are usually extracted by developing some common or special rules based on linguistic principles. Isozaki generated and refined rules by decision tree learning, overcoming the problem that hand-crafted rules were difficult to maintain [5]. By applying the refined rules, They got named entity candidates. Then non-overlapping candidates were selected by a kind of longest match method. Chen and Ma employed statistical and morphological rules to extract Chinese new words [6]. However, they only considered two-character new words. Meng et al. tackled the problem of low-frequency new words. They used parsing information to extract new words and the rules were built on the sentences [7]. For the rule-based methods, defining rules is difficult and the rules are often domain-specific which result in poor flexibility and portability.

Statistical methods usually employ statistical linguistic features, or combine with machine learning

methods to extract new words. Statistical features can be measured by pointwise mutual information (PMI) [11], independent word probability (IWP), word formation power (WFP), enhanced mutual information (EMI) [12] and multi-word expression distance (MED) [13]. Luo and Sun firstly analyzed nine statistical internal measures, then tried to improve the performance by properly combining these measures [14]. Finally, the contextual measure was integrated with internal measures to acquire more improvement. Experimental results showed the combination of internal measures and contextual measures achieved the best performance. However, they only focused on two-character Chinese words. In [9], all potential unknown words were classified into single-character and affix model based on structures of unknown words. Then some filtration methods based on statistical information were performed. Because it is difficult for these methods to choose the appropriate threshold, some researchers try to combine machine learning strategies, such as support vector machine (SVM) [15–18], conditional random fields (CRF) [19,20], hidden Markov model (HMM) [21–23] and so on. Peng et al. [8] and Xu et al [24] used the model of CRF. Peng et al. considered word segmentation and new word extraction as a unified process. They employed CRF to perform word segmentation. He and Zhu proposed a bootstrap method to extract new words [10]. Mutual information and entropy [25–27] were used. In [28], they treated new word detection as a binary classification problem and used SVM to identify Chinese new words. Zhou proposed a discriminative Markov model to detect new words by chunking one or more separated words [29]. The statistical methods need training on large scale corpus, thus manually annotating is time-consuming. Besides, it may produce sparse data, which will lead to low accuracy. Our work consider domain specificity and combine with multiple statistical language knowledge, which can overcome their disadvantages.

Hybrid methods are the combination of rule-based methods and statistical methods. Huang et al. designed statistical measures to quantify the utility of lexical patterns and the extracted patterns could be further used in finding new words [4]. The shortcoming of this method is only extracting adjective new words. In [30], they proposed to use statistical information to provide the internal criteria, simultaneously employing rule-based methods to capture external criteria. The hybrid methods suffer from the limitation of available statistical features and rules.

3 Our approach

3.1 The overview of our approach

For new word extraction, our idea is inspired by the domain specificity. New words in SD rarely appear in a different domain named as RD. Therefore, we collect relevant contents from SD, e.g., Baidu Tieba corpus, which is a good data source for extracting new network words. After the preprocessing, we firstly extract bi-grams, tri-grams, four-grams and five-grams. Then, a filtering algorithm based on domain specificity is performed to obtain a candidate list of new words. Finally, we introduce a ranking method for new word extraction, based on statistical language knowledge like string frequency, string cohesion and string liberalization. In the following subsection, we present the above steps in detail.

3.2 A filtering algorithm based on domain specificity

Firstly we extract bi-grams, tri-grams, four-grams and five-grams from SD. After filtering the n-grams containing stop words, there are many normal words and garbage strings, such as "学校 (school)", "高跟鞋 (heels)" and "要考[1)]", which are noise words for our new word extraction. It is intuitive that new words rarely appear in a different domain named as RD, but the normal words and garbage strings are common phenomenon in the n-grams of RD. Therefore, we introduce a filtering algorithm based on domain specificity, which is given in Algorithm 1.

In Algorithm 1, firstly, the corpus of RD is split into sentences according to the punctuation marks. Then, we extract n-grams from RD as set RG. If our n-gram extracted from SD occurs in RG, the gram

1) It is a garbage string and has no meaning.

Algorithm 1 A filtering algorithm based on domain specificity

Input: All the n-grams ($n \in \{2,3,4,5\}$) extracted from SD as set SG, the corpus of RD;
Output: A candidate list of new words CW;
1: Extract n-grams($n \in \{2,3,4,5\}$) from RD as set RG;
2: $CW = \Phi$;
3: **for** each n-gram named g in set SG **do**
4: **if** $g \in$ RG **then**
5: SG = SG $- g$;
6: **else**
7: CW = CW $+ g$;
8: **end if**
9: **end for**
10: **return** CW

Table 2 Examples of high string cohesion and low string cohesion

String	High cohesion	String	Low cohesion
楼主 a)	0.54	楼们 b)	0.07
思密达 a)	0.47	我个一 b)	0.09
萌妹纸 a)	0.42	我吧你 b)	0.08
喜大普奔 a)	0.39	为你比我 b)	0.14
深藏功与名 a)	0.41	我不就是想 b)	0.09

a) New word with high cohesion; b) low-cohesion string with no meaning.

will be neglected. Otherwise, it will be added into the candidate list of new words. Finally, we get a candidate list of new words.

3.3 Measuring word features

After the filtering algorithm based on domain specificity, there are many noise candidate new words, such as "富帅不[2)]", "密达我 [2)]" and "大普奔 [2)]". Therefore, we should consider the special features of words. Our approach takes use of three kinds of statistical language knowledge to measure word features.

3.3.1 *String frequency*

It is intuitive that a string can be potential word if it has high frequency. For example, the frequency of "高富帅 (a tall, rich and handsome man)" is generally high than the frequency of "富帅不 [2)]". When $S = s_1 s_2 \cdots s_n$ expresses a string, string frequency F is the number of S occuring in the corpus.

3.3.2 *String cohesion*

String cohesion is the correlation of different components in $S = s_1 s_2 \cdots s_n$, which indicates word internal feature. If a string can be a potential word, it must have strong cohesion. For example, "思密达 (new modal particle)" is a better potential word than "密达我 [2)]" due to the strong cohesion. Enhanced mutual information [12] is a useful criterion to evaluate string cohesion, which is defined as the ratio of its probability of being a multi-character to its probability of not being a multi-character:

$$C(S) = \log_2 \frac{F(S)}{\prod_{i=1}^{n}(F(s_i) - F(S))}, \tag{1}$$

where $F(S)$ and $F(s_i)$ respectively denote the string frequency of S and s_i. The key idea of string cohesion is to measure a string's dependency of internal feature. The larger the value is, the more possible the expression will be a potential new word. Table 2 gives some examples of high string cohesion and low string cohesion.

2) It is a garbage string and has no meaning.

Table 3 Examples of left entropy and right entropy

String	High liberalization		String	Low liberalization	
	Left	Right		Left	Right
蛇精病 a)	0.28	0.65	精病 b)	0.03	0.62
脑残粉 a)	0.60	0.89	残粉 b)	0.002	0.86
韩国棒子 a)	0.47	0.70	韩国棒 b)	0.42	0.009
羡慕嫉妒恨 a)	0.38	0.46	慕嫉妒 b)	0	0.06
哭晕在厕所 a)	0.61	0.37	晕在厕 b)	0.06	0

a) New word with high liberalization; b) low-liberalization string with no meaning.

3.3.3 *String liberalization*

String liberalization indicates neighborhood feature of a string. If a string can be a potential word, it will be more commonly used with diversified neighborhood. That is to say, the word has high liberalization and can be used in many different linguistic scenarios. For example, "喜大普奔 (good news, and want to tell others)" is a better potential word than "大普奔 2)" due to the high string liberalization. This can be measured by information entropy, which is usually used to indicate the degree of uncertainty or randomness. If all the left characters of S are set $C_l = \{c_1, c_2, \ldots, c_l\}$ and all the right characters of S are set $C_r = \{c_1, c_2, \ldots, c_r\}$, the left entropy and right entropy of S are as follows:

$$L_l(S) = -\sum_{i=1}^{l} \frac{F(c_i S)}{F(C_l S)} \times \log \frac{F(c_i S)}{F(C_l S)}, \tag{2}$$

$$L_r(S) = -\sum_{i=1}^{r} \frac{F(S c_i)}{F(S C_r)} \times \log \frac{F(S c_i)}{F(S C_r)}, \tag{3}$$

where $F(c_i S)$ and $F(S c_i)$ respectively denote the string frequency of $c_i S$ and $S c_i$. $F(C_l S)$ is the sum of $F(c_i S)$ $(c_i \in C_l)$. It is the same with $F(S C_r)$. The key idea of string liberalization is to measure a string's diversity of neighborhood features. If the left or right neighbor of a string is contributed by a few fixed characters, the entropy will be low and the string liberalization is also very low. Table 3 gives some examples of left entropy and right entropy.

3.4 A ranking method for new word extraction

Considering the aforementioned features in Subsection 3.3, we combine string frequency, string cohesion and string liberalization together for new word extraction. Thus, the possibility of new words can be formulated as follows:

$$\mathrm{FCL}(w) = \alpha \hat{F}(w) + \beta \hat{C}(w) + \gamma \log \frac{\hat{L}_l(w) + \hat{L}_r(w)}{|\hat{L}_l(w) - \hat{L}_r(w)|^{\sigma}}, \tag{4}$$

where $\alpha + \beta + \gamma = 1$. $\hat{F}(w)$ and $\hat{C}(w)$ are the normalized forms of string frequency and string cohesion respectively. $\hat{L}_l(w)$ and $\hat{L}_r(w)$ are the normalized forms of left entropy and right entropy. $\alpha, \beta, \gamma \in [0, 1]$ determines which type of features dominates new word extraction. $\alpha = 0$ means the possibility of candidate new words is estimated by only considering string cohesion and string liberalization, ignoring the contribution of string frequency. Otherwise, when $\alpha = 1$, the possibility of candidate new words is estimated by only considering string frequency. β, γ have the similar roles with α. σ is used to adjust the impact of the difference by left entropy and right entropy.

After computing the FCL values of the candidate new words, we obtain a ranked list of all the words. In order to finish the new word extraction, we set a threshold K. The top K new words will be considered to be the new words. We denote our proposed method as FCL.

4 Experiments and discussion

In this section, to evaluate our method, we conduct the following experiments: (1) we compare our method to several baselines; (2) we compare the impact of different features on new network word extraction; (3) we perform parameter tuning with extensive experiments; (4) we demonstrate how new network words benefit word segmentation; and (5) we demonstrate our approach domain-independent and language-independent.

4.1 Experiment setup

4.1.1 *Datasets*

To evaluate our method, we used four datasets. (1) Baidu Tieba is a good social website data source for extracting new network words. Therefore, we crawled 3524584 Tieba posts [3] to evaluate our proposed method. These posts range from January to December of 2014. (2) The News corpus is provided by the NIST open machine translation evaluation (OpenMT 2015) [31,32], which contains 9517292 Chinese sentences and 1744025 English documents of xinhuanet [4]. (3) We crawled 10237813 Weibo posts [5], applied in the word segmentation task. (4) In the domain word extraction task, we crawled papers from the ACM website [6], which consists of 2924 full papers of four conferences (CIKM, SIGIR, SIGKDD and WWW), ranging from the year 2011 to the year 2013. The Weibo posts were segmented into single words using a Chinese word segmentation tool ICTCLAS [33].

The extracted new words were manually annotated, where three annotators were involved. Two annotators were requested to judge whether the extracted word was a new network word. When conflicts occurred, the third annotator made final judgement. The annotation led to 1193 new network words and 1730 domain words.

4.1.2 *Evaluation metrics*

We select precision (P), recall (R), f-measure (F) as metrics:

$$P = \frac{\text{number of correct extraction}}{\text{total number of extraction}},$$

$$R = \frac{\text{number of correct extraction}}{\text{total number of new network words}},$$

$$F = (2 \times P \times R)/(P + R).$$

4.2 Our method vs. different baselines

To prove the effectiveness of the proposed FCL method, we select some methods for comparison as follows: new word detection (NWD) method [4], mutual information (MI) method used in [10] and pointwise mutual information (PMI) method [11]. For the MI and PMI methods, we add string frequency and string liberalization into them and set the same parameters for the fairness of comparison. The source domain we use is Tieba posts and the reference domain is Chinese News corpus. Figure 1 presents experimental results.

Observing from Figure 1, we can see that our method outperforms baseline methods both in precision and recall. It proves the effectiveness of the proposed method. For the NWD method, the performance is extremely dreadful, which is mainly because this method only extracts adjective words, ignoring other parts of speech. Besides, the segmentation tool cannot perform well on the informal texts. Our method has much better performance than PMI and MI methods. We believe the reason is that EMI can measure a string's dependency of internal feature better than PMI and MI. We also notice that when K grows

3) http://tieba.baidu.com/.
4) http://www.xinhuanet.com/.
5) http://weibo.com/.
6) http://dl.acm.org/.

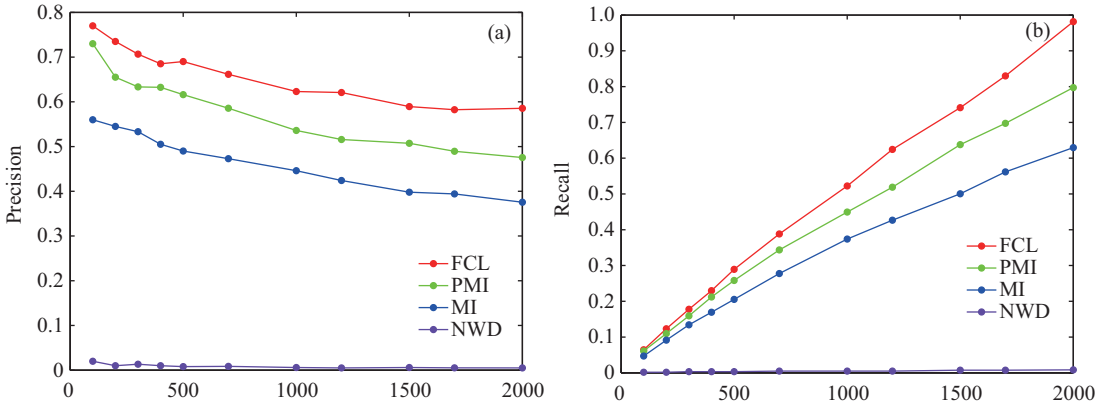

Figure 1 (Color online) (a) Precisions and (b) recalls of NWD, MI, PMI and FCL on new network word extraction. X-axis is the top words threshold K, and Y-axis is precision or recall.

Table 4 Results of different combinations of word features

Methods	$K=100$			$K=300$			$K=500$			$K=1000$			$K=1500$			$K=2000$		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
F	0.15	0.01	0.02	0.18	0.05	0.07	0.19	0.08	0.11	0.17	0.14	0.16	0.16	0.20	0.18	0.15	0.26	0.19
C	0.10	0.01	0.02	0.09	0.02	0.04	0.10	0.04	0.06	0.08	0.07	0.07	0.08	0.10	0.09	0.08	0.13	0.10
L	0.44	0.04	0.07	0.37	0.09	0.15	0.38	0.16	0.23	0.36	0.30	0.33	0.34	0.42	0.37	0.30	0.49	0.37
F+C	0.25	0.02	0.04	0.18	0.04	0.07	0.18	0.08	0.11	0.18	0.15	0.16	0.16	0.20	0.18	0.16	0.28	0.21
F+L	0.56	0.05	0.09	0.53	0.13	0.21	0.49	0.21	0.29	0.45	0.38	0.41	0.40	0.50	0.44	0.38	0.63	0.47
C+L	0.70	0.06	0.11	0.59	0.15	0.24	0.57	0.24	0.33	0.50	0.42	0.45	0.45	0.56	0.50	0.41	0.68	0.51
F+C+L	**0.77**	**0.06**	**0.12**	**0.71**	**0.18**	**0.28**	**0.69**	**0.29**	**0.41**	**0.62**	**0.52**	**0.57**	**0.59**	**0.74**	**0.66**	**0.59**	**0.98**	**0.73**

bigger, the precision decreases while the recall increases. That is because the bigger value of K can generate a wilder coverage, but bring in more noisy words. Clearly, when $K = 2000$, our method could cover 98% of the whole new network words set. Thus, it demonstrates our method is quite effective in extracting a large number of new network words.

4.3 Evaluation of different statistical language knowledge

In this subsection, we discuss which combination of word features is more effective for new network word extraction. For comparison, we design six baselines, noted as F, C, L, F+C, F+L and C+L. F only employs string frequency. C only employs string cohesion. L only employs string liberalization. F+C considers both string frequency and string cohesion. F+L and C+L have the similar way with F+C. Moreover, F+C+L is our method which considers all word features, referring to Eq. (4) with $\alpha = 0.3, \beta = 0.4, \gamma = 0.3$ and $\sigma = 0.1$. Table 4 presents experimental results when the top words threshold K varies.

From Table 4, we observe that F+C, F+L and C+L perform better than F and C. F+L and C+L perform better than L. These results indicate every word feature is necessary for new network word extraction. Moreover, F+C+L notably outperforms other baselines in all different K. It demonstrates combination of all different word features is effective.

4.4 Parameter tuning

In this subsection, we discuss the variation of extraction performance when changing α, β, γ and σ in Eq. (4). Mean reciprocal rank (MRR) is an effective metric to test the extraction performance. Therefore, we select a list of annotated new words with size of n, named DIC. MRR is computed by $\mathrm{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}$, where p_i is the ranking position of every new word in DIC. Experimental results are shown in Table 5 and Figure 2.

Table 5 The MRR-values of new network word extraction with varying α, β, γ

α	β	γ	MRR	α	β	γ	MRR
0.8	0.1	0.1	0.0050	0.1	0.1	0.8	0.0109
0.5	0.3	0.2	0.0119	0.2	0.3	0.5	0.0195
0.2	0.5	0.3	0.0198	0.4	0.4	0.2	0.0186
0.1	0.8	0.1	0.0026	0.3	0.3	0.4	0.0205
0.3	0.5	0.2	0.0152	**0.3**	**0.4**	**0.3**	**0.0218**
0.5	0.2	0.3	0.0178	0.4	0.3	0.3	0.0211

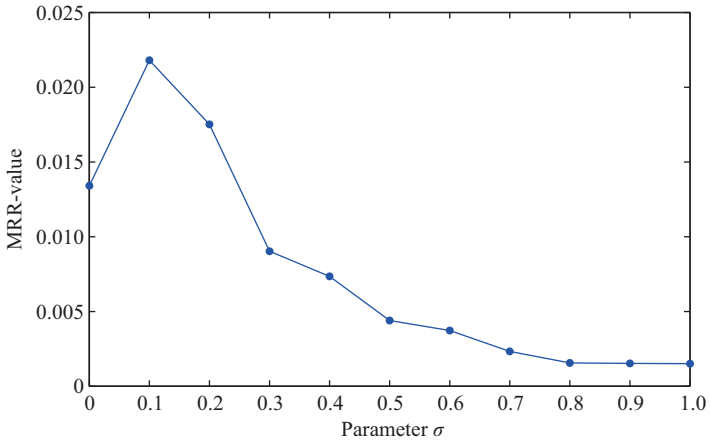

Figure 2 (Color online) The MRR-values of new network word extraction with varying parameter σ.

Table 6 The accuracy of word segmentation with four lexicons

Lexicon	Accuracy(%)	Lexicon	Accuracy(%)
DL	77.89	DL+CNW	89.17
DL+T1000	87.54	DL+ANW	89.60

Table 5 presents the MRR-values of new network word extraction with varying α, β, γ while fixing $\sigma = 0.1$. Due to the space limitation, we only show twelve groups of parameter settings. We observe the best performance is obtained when $\alpha = 0.3, \beta = 0.4, \gamma = 0.3$. It indicates that string frequency, string cohesion and string liberalization are all useful for new network word extraction. The performance benefits from their combination. Figure 2 present the MRR-values with varying σ from 0 to 1 while fixing $\alpha = 0.3, \beta = 0.4, \gamma = 0.3$. We notice the performance increases when σ is set from 0 to 0.1. When σ gets bigger, performance, however, decreases. The best performance is achieved when $\sigma = 0.1$.

4.5 Application of new network words to word segmentation

In this subsection, we demonstrate whether new network words would benefit Chinese word segmentation. For this purpose, we randomly sampled 500 Weibo posts that contain at least one of our annotated new network words. These posts were manually segmented into single words as ground truth. We compare four different kinds of lexicons for word segmentation. One of the lexicons is the default lexicon (DL) in ICTCLAS [33]. Moreover, we add three resources into the default lexicon: the top 1000 words produced by our approach (denoted by T1000), all correct new words produced by our approach (denoted by CNW, including 1172 new words) and all annotated new words (denoted by ANW, including 1193 new words), respectively. Thus, the four different kinds of lexicons are DL, DL+T1000, DL+CNW and DL+ANW. We use segmentation accuracy to evaluate the contribution of four lexicons to word segmentation.

The results are shown in Table 6. We can find that all three lexicons which contain new words improve the performance remarkably. The lexicon DL+T1000 is generated by our method, which increases the performance of word segmentation by 10%. The lexicon DL+CNW outperforms DL+T1000, which is mainly because T1000 may contain words that are not new words. We also observe that DL+ANW

Table 7 The MRR-values of domain word extraction with varying α, β, γ

α	β	γ	MRR	α	β	γ	MRR
0.8	0.1	0.1	0.0141	0.1	0.1	0.8	0.0200
0.5	0.3	0.2	0.0189	0.2	0.3	0.5	0.0195
0.2	0.5	0.3	0.0202	0.4	0.4	0.2	0.0178
0.1	0.8	0.1	0.0007	0.3	0.3	0.4	0.0206
0.3	0.5	0.2	0.0195	0.3	0.4	0.3	0.0207
0.5	0.2	0.3	0.0180	**0.4**	**0.3**	**0.3**	**0.0213**

Table 8 The MRR-values of domain word extraction with varying parameter σ

σ	MRR	σ	MRR	σ	MRR
0.1	0.0139	**0.2**	**0.0206**	0.3	0.0187
0.4	0.0184	0.5	0.0181	0.6	0.0180
0.7	0.0184	0.8	0.0183	0.9	0.0187

outperforms DL+CNW. The reason is that the number of annotated new words is bigger than the number of correct new words produced by our approach.

4.6 Domain word extraction in English academic area

In this subsection, we demonstrate our approach domain-independent and language-independent. We try to employ our approach to extract English academic words of computer science, such as "machine learning", "topic model" and "natural language processing". The source domain we use is ACM papers and the reference domain is English News corpus.

4.6.1 *Experiment setup*

To evaluate the variation of extraction performance in English corpus when changing α, β, γ and σ, we compute MRR-values separately with different parameters. Table 7 presents the MRR-values of domain word extraction with varying α, β, γ while fixing $\sigma = 0.2$. We observe the best performance is obtained when $\alpha = 0.4, \beta = 0.3, \gamma = 0.3$. Table 8 present the MRR-values with varying σ from 0.1 to 0.9 while fixing $\alpha = 0.4, \beta = 0.3, \gamma = 0.3$. We notice the best performance is achieved when $\sigma = 0.2$. Finally, we set $\alpha = 0.4, \beta = 0.3, \gamma = 0.3$ and $\sigma = 0.2$.

4.6.2 *Extraction of domain words*

To prove the effectiveness of our method, we select the two baselines: MI method and PMI method, as described in Subsection 4.2. The NWD method is not involved because it only extracts adjective new words, which is meaningless to domain word extraction in English academic area. Figure 3 presents experimental results.

From Figure 3, we observe that our method outperforms baseline methods both in precision and recall. It proves the effectiveness of the proposed method on domain word extraction. We also notice that when K grows bigger, the precision decreases while the recall increases, which is the same reason with new network word extraction experiment in Subsection 4.2. The best precision is obtained when $K = 100$ and the best recall is obtained when $K = 2000$. Experimental results show our method could extract 1634 domain words when $K = 2000$, which demonstrates our method is very effective in recognizing a great number of domain words.

5 Conclusion and future work

In this paper, we propose a novel approach to the task of new word extraction. The approach is fully unsupervised, purely data-driven and adaptive in multiple languages. We perform a filtering algorithm based on domain specificity and employ the statistical language knowledge considering word frequency

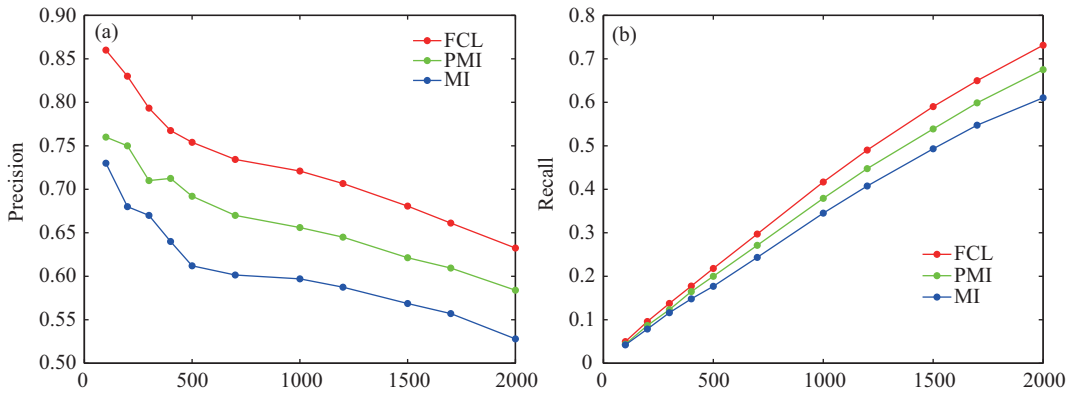

Figure 3 (Color online) (a) Precisions and (b) recalls of FCL, PMI and MI on domain word extraction. X-axis is the top words threshold K, and Y-axis is precision or recall.

feature, word internal feature and neighborhood feature to extract new words. The proposed method does not need the rules defined by human to filter undesirable words. Compared to different baselines, experimental results prove the effectiveness of our method both in Chinese and English corpus. What's more, experiments also demonstrate new network words benefit word segmentation obviously.

The pure statistical method usually need to adjust the weights of combination and threshold. For the future work, we intend to combine our method with some known methods, e.g. genetic algorithm, to adjust the weights. We are also considering how to excavate more useful features to improve the performance of new word extraction.

Acknowledgements This work was supported by State Key Program of National Natural Science of China (Grant No. 61132009), National High Technology Research and Development Program of China (863 Program) (Grant No. 2015AA015404) and National Natural Science Foundation of China (Grant Nos. 61201351, 61402036).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Sproat R, Emerson T. The first international Chinese word segmentation bakeoff. In: Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2003. 17: 133–143
2 Sun X, Wang H, Li W. Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012. 1: 253–262
3 Nie L, Yan S, Wang M, et al. Harvesting visual concepts for image search with complex queries. In: Proceedings of the 20th ACM International Conference on Multimedia. New York: ACM, 2012. 59–68
4 Huang M, Ye B, Wang Y, et al. New word detection for sentiment analysis. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, 2014. 531–541
5 Isozaki H. Japanese named entity recognition based on a simple rule generator and decision tree learning. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2001. 314–321
6 Chen K J, Ma W Y. Unknown word extraction for Chinese documents. In: Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002. 1: 1–7
7 Meng Y, Yu H, Nishino F. Chinese new word identification based on character parsing model. In: Proceedings of the 1st International Joint Conference on Natural Language Processing, Hainan, 2004. 489–496
8 Peng F, Feng F, McCallum A. Chinese segmentation and new word detection using conditional random fields. In: Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2004. 562
9 Jiang X, Wang L, Cao Y, et al. Automatic recognition of Chinese unknown word for single-character and affix models. In: Knowledge Engineering and Management. Berlin: Springer, 2011. 435–444
10 He S, Zhu J. Bootstrap method for Chinese new words extraction. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, 2001. 1: 581–584
11 Church K W, Hanks P. Word association norms, mutual information, and lexicography. Comput Linguist, 1990, 16: 22–29

12 Zhang W, Yoshida T, Tang X, et al. Improving effectiveness of mutual information for substantival multiword expression extraction. Expert Syst Appl, 2009, 36: 10919–10930
13 Bu F, Zhu X, Li M. Measuring the non-compositionality of multiword expressions. In: Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010. 116–124
14 Luo S, Sun M. Two-character Chinese word extraction based on hybrid of internal and contextual measures. In: Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2003. 17: 24–30
15 Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. New York: ACM, 1992. 144–152
16 Qu A P, Chen J M, Wang L W, et al. Segmentation of Hematoxylin-Eosin stained breast cancer histopathological images based on pixel-wise SVM classifier. Sci China Inf Sci, 2015, 58: 092105
17 Zou B, Peng Z M, Xu Z B. The learning performance of support vector machine classification based on Markov sampling. Sci China Inf Sci, 2013, 56: 032110
18 Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Tech (TIST), 2011, 2: 27
19 Lafferty J, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc, 2001. 282–289
20 Yi J, Peng Y X, Xiao J G. A temporal context model for boosting video annotation. Sci China Inf Sci, 2013, 56: 110904
21 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE, 1989, 77: 257–286
22 Suk M, Ramadass A, Jin Y, et al. Video human motion recognition using a knowledge-based hybrid method based on a hidden Markov model. ACM Trans Intell Syst Tech, 2012, 3: 42
23 Hong F, Tang J W, Lu P P. Multichannel DEM reconstruction method based on Markov random fields for bistatic SAR. Sci China Inf Sci, 2015, 58: 062302
24 Xu Y S, Wang X, Tang B Z, et al. Chinese unknown word recognition using improved conditional random fields. In: Proceedings of the 8th International Conference on Intelligent Systems Design and Applications, Kaohsiung, 2008. 2: 363–367
25 Hu Q H, Guo M Z, Yu D R, et al. Information entropy for ordinal classification. Sci China Inf Sci, 2010, 53: 1188–1200
26 Sun Y L, Tao J X, Chen H, et al. The entropy weighted non-uniform scanning algorithm for diffraction tomography. Sci China Inf Sci, 2015, 58: 067102
27 Ding Y, Zhang Y, Wang X, et al. Perceptual image quality assessment metric using mutual information of Gabor features. Sci China Inf Sci, 2014, 57: 032111
28 Li H, Huang C N, Gao J, et al. The use of SVM for Chinese new word identification. In: Natural Language Processing—IJCNLP 2004. Berlin: Springer, 2005. 723–732
29 Zhou G D. A chunking strategy towards unknown word detection in Chinese word segmentation. In: Proceedings of the 1st International Joint Conference on Natural Language Processing. Berlin: Springer, 2005. 530–541
30 Wu A D, Jiang Z X. Statistically-enhanced new word identification in a rule-based Chinese system. In: Proceedings of the 2nd Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2000. 12: 46–51
31 Liberman M, Davis K, Grossman M, et al. Emotional Prosody Speech and Transcripts. LDC2002S28. Philadelphia: Linguistic Data Consortium, 2002
32 Huang S D, Graff D, Doddington G. Multiple-Translation Chinese Corpus. LDC2002T01. Philadelphia: Linguistic Data Consortium, 2002
33 Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS. In: Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Stroudsburg: Association for Computational Linguistics, 2003. 17: 184–187