Unknown Word Recognition Based on Extended Rules and Statistical Features (Postprint)

Authors: Zeng Hao, ZHAN Enqi, Zheng Jianbin, Wang Yang

Date: 2018-05-24T00:00:00+00:00

Abstract

To improve the recognition performance of out-of-vocabulary (OOV) words across various industry domains, this paper proposes an OOV word recognition method based on expansion rules and statistical features. The method analyzes the morphological characteristics of OOV words in industry domains to formulate expansion rules. According to these rules, segmentation units are expanded to obtain compound words, which are then evaluated using statistical features such as word frequency, mutual information, and neighbor entropy to determine whether they are OOV words. If identified as OOV words, they undergo further expansion and recognition. Experimental results on OOV word recognition in six industry domains and a general domain demonstrate that the proposed method achieves favorable recognition performance and exhibits good portability.

Full Text

Preamble

Unregistered Word Recognition Based on Extended Rules and Statistical Features

Zeng Hao[1,2], *Zhan Enqi*[1,2], *Zheng Jianbin*[1,2], *Wang Yang*[1,2]†
[1]School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China
[2]Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Ministry of Education, Wuhan 430070, China

Computer Application Research (ChinaXiv Cooperative Journal)
http://www.arocmag.com/article/02-2019-09-007.html

Abstract

To improve unregistered word recognition performance across various industry domains, this paper proposes a method based on expansion rules and statistical features. The approach analyzes the morphological characteristics of domain-specific unregistered words to formulate expansion rules, which are then applied to segmentation units to generate compound words. Statistical features including word frequency, mutual information, and branch entropy are employed to determine whether a compound word qualifies as an unregistered word, with further expansion and recognition performed if it does. Experimental results from six industry domains and a general domain demonstrate that the proposed method achieves favorable recognition performance and exhibits strong portability.

Keywords: unregistered word; expansion rules; word frequency; mutual information; branch entropy

1 Related Research

Unregistered words refer to terms not included in segmentation dictionaries and newly emerging words that appear with changing times. Recognition methods can be categorized into rule-based approaches, statistical approaches, and hybrid approaches combining rules and statistics.

In written English and other Western languages, punctuation marks separate sentences while spaces separate words, enabling straightforward word identification. In contrast, while Chinese also uses punctuation for sentence boundaries, words lack explicit delimiters—characters appear continuously, with any adjacent characters potentially forming a word of unrestricted length. Consequently, Chinese word segmentation becomes a critical foundational task for computer processing of Chinese text. Although mainstream segmenters achieve high accuracy in general domains, their performance in specialized industry domains remains unsatisfactory. This is primarily because domain-specific unregistered words are typically longer, semantically complete compound words containing special characters, making them difficult to recognize. Low recognition accuracy for domain-specific unregistered words directly impedes overall segmentation accuracy, motivating this research on industry domain unregistered word recognition.

Rule-based methods identify unregistered words through morphological patterns, part-of-speech rules, and word formation probabilities. Zheng et al. [1] studied Chinese word formation principles and established construction rules for recognizing web neologisms, achieving 91.2% accuracy. Cui et al. [2] built garbage and affix dictionaries from corpora, combining part-of-speech rules with independent word formation probabilities to detect web neologisms with good results. While rule-based methods offer high precision, their rules typically originate from specific domains, resulting in poor portability and inability to cover all morphological phenomena.

Statistical approaches treat words as independent units that should exhibit stable internal structure and rich contextual environments, commonly employing word frequency, mutual information, and branch entropy as recognition features. Han et al. [3] used mutual information to extract bigrams and branch entropy to determine boundaries, iteratively expanding to identify longer, more semantically complete unregistered words. Yang et al. [4] integrated multiple statistical features to extract strings of length up to 6, selecting those exceeding all threshold values as unregistered words. Li et al. [5] filtered bigrams with stable structure using mutual information, then expanded them using branch entropy to identify unregistered words of length 2-4. Yao et al. [6] employed improved mutual information to obtain stable 2-grams and 3-grams, recognizing unregistered words through branch entropy calculation and 3-gram expansion. Duan et al. [7] screened candidates within certain ranges using word frequency, document frequency, and average word frequency. Pang et al. [8] analyzed word distribution characteristics across documents, within documents, and within paragraphs. Zhang et al. [9] clustered microblogs using K-means, extracted candidate strings exceeding frequency thresholds from each cluster, and used adjacency degree to determine whether substrings were unregistered words. Statistical methods offer good portability without rule dependency but suffer from high computational cost and produce numerous non-word strings due to lack of rule constraints.

To overcome these limitations, researchers increasingly favor hybrid approaches combining rules and statistics. Liu et al. [10] integrated statistical methods, domain dictionaries, part-of-speech rules, and affix rules. Huo et al. [11] combined word frequency with morphological rules. Zhou et al. [12] integrated word frequency, part-of-speech rules, and adjacency variation counts. Du et al. [13] used improved mutual information to screen bigrams for expansion, filtering results through word frequency and stopword rules.

Most existing research focuses on news and microblog corpora for general domain unregistered word recognition, primarily targeting Chinese words of length 2-4, with limited investigation into longer, semantically complete compound words. Additionally, recognition of special unregistered words containing English characters in Chinese text remains relatively understudied.

2 Industry Domain Unregistered Word Recognition

This study investigates industry domain unregistered word recognition by crawling job postings from recruitment websites to build a position corpus. Job postings typically consist of structured data (salary, location, publication date) and unstructured data (job responsibilities, requirements, benefits). Since position information concentrates in the unstructured portion, subsequent processing focuses exclusively on this component.

Fifty positions were extracted from each of six industry domains and segmented using the HanLP segmenter. As segmentation errors primarily stem from am-

biguity and unregistered words, fields resulting from ambiguity were excluded, leaving remaining error fields as unregistered words. Table 2 presents the unregistered word statistics across six industry domains.

Unregistered words in Table 2 can be categorized as Chinese or English. Chinese unregistered words account for approximately 90%, primarily comprising personal names, place names, organization names, and technical terminology. English unregistered words constitute about 10%, mainly representing job skills such as "c++" and "j2se" . Tables 3 and 4 analyze the morphological characteristics of Chinese and English unregistered words, respectively.

Table 3 reveals that Chinese domain unregistered words typically consist of 2-3 Chinese words forming compound words. Table 4 shows that English unregistered words also comprise 2-3 components but exhibit more flexible formation patterns. While Chinese words primarily combine with other Chinese words, English words can combine with Chinese words (e.g., "c language"), numbers (e.g., "html5"), or special characters (e.g., "c#"). HanLP fails to recognize these unregistered words, incorrectly segmenting them into multiple units—for instance, splitting "深度学习" into "深度/学习" and "j2ee" into "j/2/ee" . Therefore, by leveraging morphological characteristics to recombine segmentation units according to specific rules and applying filtering strategies, domain-specific unregistered words can be effectively recognized.

3 Unregistered Word Recognition Based on Expansion Rules and Statistical Features

3.1 Expansion Rules

Building upon the morphological analysis of domain unregistered words, this paper proposes a recognition method combining expansion rules with statistical features. "Expansion" refers to combining the current word with its subsequent word within the same sentence' s segmentation result to form a compound word. The expansion rules are defined as follows:

Rule 1: If the current word is a stopword or neither Chinese nor English, it cannot be expanded.

Rule 2: If the current word is a Chinese word and not a stopword, and the subsequent word is also a Chinese word and not a stopword, the current word can be expanded.

Rule 3: If the current word is an English word and not a stopword, and the subsequent word is not a stopword, the current word can be expanded.

Rule 4: If the expansion count exceeds the preset maximum, no further expansion is permitted.

These rules derive from summarizing domain unregistered word formation characteristics. While Chinese words typically combine only with other Chinese

words, English words can form meaningful compounds with Chinese words, numbers, and special characters. These expansion rules thus filter compounds matching domain unregistered word patterns while eliminating meaningless combinations to improve recognition effectiveness.

Expansion rules require a stopword dictionary. In natural language processing, stopwords contribute only syntactic function without semantic meaning, typically not forming meaningful compounds with other words (e.g., “了”, “的”, “不”). While various general-domain stopword dictionaries exist online, this research spans multiple industry domains. To enhance recognition performance, the position corpus was segmented and word frequencies calculated. Words with frequency exceeding 1000 and low compound formation probability were selected as domain-specific stopwords. Table 5 shows partial stopwords and their frequencies. Combined with a general-domain stopword dictionary, this yielded an industry-domain stopword dictionary containing 1,900 entries.

3.2 Statistical Features

This paper employs word frequency, mutual information, and branch entropy as statistical features for unregistered word recognition. A compound word is identified as an unregistered word only if all its statistical feature values exceed corresponding thresholds.

1) Word Frequency As words, unregistered words must appear with certain frequency. Let $f(w)$ denote the occurrence count of compound word w in the corpus. Larger $f(w)$ indicates higher probability of w being an unregistered word.

2) Mutual Information As words, unregistered words should exhibit stable internal structure. In information theory, mutual information (MI) measures the correlation between two signals and can thus assess the cohesion between two words. Higher mutual information indicates tighter binding and greater probability that adjacent words form an unregistered word. The mutual information calculation formulas are given in equations (1)-(4).

These formulas apply only to compound words consisting of two words. To calculate mutual information for multi-word compounds, equation (1) is modified as shown in equation (5).

Where $w_1, w_2, ..., w_n$ are the n words comprising compound word w, $MMI(w)$ is the improved mutual information of w, and $Avg(w)$ represents the average probability across different combinations of w. For example, for the three-word compound “自然语言处理” (natural language processing), the average probability across its different combinations is:

3) Branch Entropy As words, unregistered words should possess rich contextual environments. Branch entropy (BE) is a crucial statistical feature for measuring word formation probability, utilizing information entropy (IE) to

calculate the uncertainty of a string' s context [14]. In information theory, information entropy represents the mean uncertainty of a random variable—larger entropy indicates greater uncertainty. For a discrete random variable X with value space A, when X takes value a, the information entropy is given by equation (7).

Branch entropy comprises left branch entropy (LBE) and right branch entropy (RBE). Larger left branch entropy indicates richer left context and more determinate left boundaries; larger right branch entropy indicates richer right context and more determinate right boundaries. When both left and right branch entropies are substantial, the probability of the string forming an independent word increases. In this paper, a compound word' s left neighbor refers to its preceding word, and its right neighbor refers to its following word. All left neighbors constitute the left neighbor set, and all right neighbors constitute the right neighbor set. Distinct left neighbors form left neighbor categories, while distinct right neighbors form right neighbor categories. Assuming compound word w has left neighbor categories $\{L_1, L_2, ..., L_n\}$ and right neighbor categories $\{R_1, R_2, ..., R_m\}$, its left and right branch entropies are calculated using equations (8) and (9), respectively.

Where $LBE(w)$ denotes w' s left branch entropy, n is the left neighbor set size, and n_i represents the occurrence count of left neighbor L_i. In equation (9), $RBE(w)$ denotes w' s right branch entropy, m is the right neighbor set size, and m_i represents the occurrence count of right neighbor R_i.

3.3 Unregistered Word Recognition Process

Figure 2 [Figure 2: see original paper] illustrates the unregistered word recognition process based on expansion rules and statistical features. The specific steps are:

a) Set maximum expansion count, word frequency threshold, mutual information threshold, left branch entropy threshold, and right branch entropy threshold.

b) Split the corpus into short sentences using Chinese punctuation marks.

c) Segment short sentences using HanLP and traverse segmentation units. Apply expansion rules to determine if the current word is expandable. If not, skip it and treat the next segmentation unit as the current word for expansion and recognition. If expandable, calculate the compound word' s statistical feature values. If all values exceed thresholds, add to the unregistered word set and continue expanding and recognizing the compound word; otherwise, discard the compound word and treat the next segmentation unit as the current word.

d) After processing all short sentences in the corpus, the algorithm terminates and outputs the unregistered word set.

4 Experiments and Analysis

4.1 Experimental Methods and Evaluation Criteria

Parameter settings for the proposed method: maximum expansion count of 2 (recognizing only two-word and three-word compounds), word frequency threshold of 10, mutual information threshold of 3, and left/right branch entropy thresholds of 1. To account for low-frequency unregistered words in the corpus, these thresholds are set relatively low to maximize recognition coverage.

Reference [4] originally extracted strings of length 2-6, identifying unregistered words when word frequency, mutual information, left branch entropy, and right branch entropy all exceeded thresholds. For comparison with our method targeting longer, semantically complete compounds, we modified this approach to extract bigrams and trigrams from adjacent segmentation units rather than fixed-length strings, maintaining identical threshold values.

Reference [5] originally extracted two-character combinations from fragmented strings using mutual information and word frequency thresholds, then expanded them using left/right branch entropy to identify unregistered words of length 2-4. To enable comparison with our compound-focused method, we modified it to extract bigrams from adjacent segmentation units while preserving the original expansion and recognition process and threshold values.

Evaluation metrics include precision (P), recall (R), and F-score (F), calculated as shown in equations (10)-(12).

Where C represents the set of unregistered words identified by the method, and D represents the manually annotated unregistered word set.

4.2 Industry Domain Unregistered Word Recognition Comparison

Applying the methods from references [4] and [5] and our proposed method to identify unregistered words in 50 positions from each of the six industry domains in Table 2 , the experimental results are presented in Tables 6 through 8 .

The results demonstrate that our method achieves superior performance in industry domain unregistered word recognition, with higher precision, recall, and F-score compared to the other two methods. While references [4] and [5] focus on general domain recognition using microblog corpora and leverage statistical features effectively, they lack rule-based constraints, resulting in numerous non-word strings exceeding statistical thresholds. For example, reference [4]' s method identified "学习 Java" (learn Java) as an unregistered word in the IT/Internet domain because "学习" and "Java" co-occurred frequently, yielding high word frequency, mutual information, and branch entropy values. Our method combines statistical features with expansion rules derived from domain unregistered word morphological analysis. Since Chinese words typically combine only with other Chinese words while English words can combine with Chinese words, numbers, and special characters, our expansion rules prevent

generation of meaningless combinations like “学习 Java” , thereby improving recognition effectiveness.

4.3 General Domain Unregistered Word Recognition Comparison

Microblogs cover broad content and represent general domain data. We selected 5,000 microblogs from the COAE2014 dataset and applied references [4] and [5] and our method to identify unregistered words. Due to the difficulty of annotating microblog unregistered words comprehensively, we evaluated results using precision only, as shown in Table 9 .

The results indicate that our method also performs well on microblog unregistered word recognition, achieving higher precision than references [4] and [5].

5 Conclusion

This paper investigates industry domain unregistered word recognition by crawling job postings from recruitment websites, analyzing morphological characteristics to formulate expansion rules, generating compound words through rule-based expansion, and evaluating them using word frequency, mutual information, and branch entropy. Experiments across six industry domains and a general domain demonstrate that our method achieves favorable precision, recall, and F-score, confirming its effectiveness and strong portability.

A limitation of our method is its strict requirement that all statistical feature values must exceed thresholds, causing failure to recognize some unregistered words where certain features fall below thresholds. For instance, words with word frequency, mutual information, and left branch entropy exceeding thresholds but right branch entropy below threshold cannot be identified. Future work will address this issue to further improve recognition performance.

References

[1] Zheng Jiaheng, Li Wenhua. A new approach to automatic recognition of web new words based on word formation [J]. Journal of Shanxi University: Natural Science Edition, 2002, 25 (2): 115-119.

[2] Cui Shiqi, Liu Qun, Meng Yao, et al. New word detection based on large-scale corpus [J]. Journal of Computer Research and Development, 2006, 43 (5): 927-932.

[3] Han Yan, Lin Yuxi, Yao Jianmin. Extended identification method for unregistered words based on statistical information [J]. Journal of Chinese Information Processing, 2009, 23 (03): 24-30, 50.

[4] Yang Yang, Liu Longfei, Wei Xianhui. Emotional new word discovery method based on word vector [J]. Journal of Shandong University: Science Edition, 2014, 49 (11): 51-58.

[5] Li Wenkun, Zhang Yangsen, Chen Ruoyu. New word discovery based on internal conjunction degree and boundary freedom [J]. Application Research of Computers, 2015, 32 (8): 2302-2304, 2342.

[6] Yao Rongpeng, Xu Guoyan, Song Jian. Microblog new word discovery method based on improved mutual information and adjacency entropy [J]. Journal of Computer Applications, 2016, 36 (10): 2772-2776.

[7] Duan Yufeng, Ju Fei. Research on recognition of chinese new words in professional field based on N-Gram [J]. Data Analysis and Knowledge Discovery, 2012, 28 (2): 41-47.

[8] Pang Wenbo, Fan Xiaozhong, Gu Yijun, et al. Chinese unknown words extraction based on word-level characteristics [C]// Proc of International Conference on Hybrid Intelligent Systems. 2009: 361-366.

[9] Zhang Shuai, Liu Qianren, Wang Lei. A Weibo-oriented method for unknown word extraction [C]// Proc of the 8th International Conference on Semantics, Knowledge and Grids. Washington DC: IEEE Computer Society, 2012: 209-212.

[10] Liu Qingtang, Wu Linjing, Yang Zongkai, et al. Domain phrase identification using atomic word formation in Chinese text [J]. Knowledge-Based Systems, 2011, 24 (8): 1254-1260.

[11] Huo Shuai, Zhang Min, Liu Yiqun. New word discovery method based on weibo content [J]. Pattern Recognition and Artificial Intelligence, 2014, 27 (2): 141-145.

[12] Zhou Chao, Yan Xin, Yu Zhengtao. Micro-blog new word recognition based on word frequency feature and adjacency change number [J]. Journal of Shandong University: Science Edition, 2015, 50 (3): 6-10.

[13] Du Liping, Li Xiaoge, Yu Gen. The improvement of chinese word segmentation based on new word discovery based on improved mutual information [J]. Journal of Peking University: Natural Science Edition, 2016, 52 (1): 35-40.

[14] Zhikov V, Takamura H, Okumura M. An efficient algorithm for unsupervised word segmentation with branching entropy and MDL [J]. Transactions of the Japanese Society for Artificial Intelligence, 2013, 28 (3): 331-338.

Note: Figure translations are in progress. See original paper for figures.