Discovery of New Words in Tax-related Fields Based on Word Vector Representation

Wei Wei[1,2], Wei Liu[1], Beibei Zhang[1], Rafał Scherer[3], Robertas Damasevicius[4]*

[1] School of Computer Science and Engineering, Xi'an University of Technology, China
[2] School of Mathematics and Computer Science, Shaanxi University of Technology, China
[3] Czestochowa University of Technology Al, Poland
[4] Department of Software Engineering, Kaunas University of Technology, Lithuania
weiwei@xaut.edu.cn, 1416165438@qq.com, bbzhang115@hotmail.com, rafal.schererg@pcz.pl, robertas.damasevicius@ktu.lt

Abstract

New words detection, as basic research in natural language processing, has gained extensive concern from academic and business communities. When the existing Chinese word segmentation technology is applied in the specific field of tax-related finance, because it cannot correctly identify new words in the field, it will have an impact on subsequent information extraction and entity recognition. Aiming at the current problems in new word discovery, it proposed a new word detection method using statistical features that are based on the inner measurement and branch entropy and then combined with word vector representation. First, perform word segmentation preprocessing on the corpus, calculate the internal cohesion degree of words through statistics of scattered string mutual information, filter out candidate two-tuples, and then filter and expand the two-tuples; next, it locks the boundaries of new words through calculate the branch entropy. Finally, expand the new vocabulary dictionary according to the cosine similarity principle of word vector representation. The unsupervised neologism discovery proposed in this paper allows for automatic growth of the neologism lexicon, experimental results on large-scale corpus verify the effectiveness of this method.

Keywords: New word discovery, Word internal combination degree, Boundary degree of freedom, Word vector representation

1 Introduction

With the increasing variety of tax data resources, the amount of data is growing rapidly, especially the rapid growth of unstructured data such as electronic bills, videos, and web pages in recent years. It is a necessary and essential point for new word detection in the area of Chinese information processing. It is directly related, to processing problems at all levels. Therefore, how to distinguish new words quickly and accurately from the corpus is the focus of research in natural language processing. New words, also known as unregistered words, these new words have new meanings and have not been included in the previous dictionaries. In this article, unregistered words [1] are equated with new words, that is, words that are not separated by word segmentation tools.

There is no unified standard definition of new words, and there is no universal law for their composition, so there is no universal method for new word identification. In this paper, through experiments, using the me-vector algorithm, and through reasonable adjustment of parameters, we can also achieve good results on the small-scale corpus.

In summary, our main contributions to this work are:

(1) For the new words in combination strings unique to the tax-related field, we use the word segmentation tool to preprocess the word segmentation, and then use the N_gram model to expand the pre-segmented word string to construct a candidate set of combined word vector mapping when similar words are found.
(2) We use statistical features such as multi-word mutual information of internal associativity and adjacency entropy of boundary degrees of freedom to calculate the closeness between word strings and the clarity of boundary features.
(3) We use the similarity-enhanced word vector to identify new words with low frequency but rich boundary degrees of freedom in the corpus, and calculate the word with the greatest similarity through the similarity measure as the expansion of the new word dictionary.

We first introduce the research background and research status of this paper, and then explain in detail the related technologies used, then the overall process and pseudocode implementation of the algorithms mentioned in this paper, and finally compare the experimental performance to demonstrate the proposed in this paper.

2 Research Status of Neologism Discovery

As the first step in the field of language processing, text disambiguation affects a variety of subsequent text analysis tasks, and neologisms, as the factor that most affects disambiguation accuracy [2-4], have always been the focus of research by scholars. Among them, neologisms are newly

*Corresponding Author: Wei Wei; E-mail: weiwei@xaut.edu.cn
DOI: 10.53106/160792642023072404010

created words or new uses of old words that have never been processed by the model, also known as unknown words. Traditional methods for new word discovery can be divided into two aspects based on strong supervision and weak supervision [5-7]. The strongly supervised identification method is mainly based on statistical knowledge of the labeled data [8].

Fu [9] et al. proposed a stable neologism recognition algorithm based on a survival rule model with reference to natural selection rules and forgetting rules and eliminated spam and pseudo-neologisms by analyzing the frequency variation of the temporal distribution of candidate word strings and the combined competitiveness in the linguistic environment. Li [10] et al. proposed a recognition system consisting of a combination of domain-specific neologism detection and word propagation for neologisms in tourism, combining statistical information on neologisms with data-driven iterative algorithms such as optimization functions and machine learning to improve the quality of sentiment dictionaries. Liu [11] et al. proposed combining pseudo-labeled data with multi-tasking to enhance the performance of word splitters by sharing network parameters. Strongly supervised discovery methods whose biggest drawback is the need to label the data and construct complex feature engineering [12] are difficult to meet production requirements in practice, and the accuracy is highly dependent on statistical analysis and training in a large-scale corpus.

Weakly supervised recognition methods have mainly used a combination of rule-based and lexical statistical information methods. Chen Fei [13] et al. proposed a series of statistical information quantities for discriminating new word boundaries and transformed the new word discovery problem into a word boundary determination problem by introducing conditional random fields (CRF). Luo [14] et al. proposed an unsupervised domain new word discovery method, which introduced LDA models into a new word recognition algorithm into algorithm to solve the online detection problem of new words. Huang [15] et al. proposed the introduction of a word quality assessment method for phrase extraction tasks into new word recognition detection. In today's high-standard annotated corpus deficit environment, weak supervision has a wider range of applications than strong supervision and is also more effective and practical [16].

3 Related Works

3.1 Internal Statistics

The degree of internal combination of a word is a measure of the degree of close combination of two Chinese characters, which is used to measure the possibility of two Chinese characters forming a word [17]. As a meaningful, independent, and applicable language component, neologisms must have a higher degree of relevance between the various elements within the neologism. The point mutual information has a good effect on the calculation of binary phrases, but it is more difficult to divide the multi-character words larger than two characters into two parts [18]. To solve this problem, the multi-character mutual information is used to disperse the binary strings. The calculation formula for multi-word mutual information is show in (1)-(3):

$$MMI(w_1...w_n) = \log_2 \frac{p(w_1...w_n)}{avg(w_1...w_n)}. \quad (1)$$

$$p(w_1 \cdots w_n) = \frac{f(w_1 \cdots w_n)}{num}. \quad (2)$$

$$avg(w_1 \cdots w_n) = \frac{1}{n-1}\sum_{i=1}^{n-1} p(w_1 \cdots w_i)p(w_i \cdots w_n). \quad (3)$$

Among them, $w_1 \ldots w_n$ represents a string of multiple characters, $p(w_1...w_n)$ is the probability of $w_1...w_n$ appearing in the corpus, and represents the average probability of different combinations of multi-character strings.

A good influence, mutual information, is used in the calculation of two-tuples, but for multi-character words larger than two words that segment two parts were a tricky problem. In the method of this article, the result of word segmentation is segmented, and each word is used as a new unit. This effectively solves the problem of the difficulty in dividing the multi-character words of mutual information. The 4-character words "winning bid amount" and "winning bid supplier" are segmented into "winning bid/amount" and "winning bid/supplier" to calculate the word frequency and mutual information of these two words respectively.

Select the two tuples whose mutual information is greater than the threshold [19-23] as the candidate two-tuples, and then perform iterative statistics by expanding to the adjacent elements of the candidate word, until the mutual information of the candidate word and the right adjacent string is less than the mutual information threshold, then stop expanding to the right. The method is as follows: Suppose the position of the "scattered string" element is $i(i = 1...n)$, C_i and C_{i+1} are two adjacent elements, if the mutual information value $MI(Ci, Ci+1)$ is greater than the threshold MI_TH, then this collocation $(Ci, Ci+1)$ is recorded as a candidate new word CNword $(Ci, Ci+1)$, and expanded to the right, When and only when the mutual information value of $Ci\&Ci+1\&Ci+2\&...\&Ci+m$ and $Ci+m+1$ is less than the threshold, the expansion stops, and $Ci\&Ci+1\&Ci+2\&... \&Ci+m$ is counted into the candidate set. Take the new term "purchasing agent mechanism" as an example. When the binary new word string "purchasing agent" is counted, if it is higher than the threshold, it will continue to expand to the right and calculate the mutual information between "purchasing agent" and "mechanism". Come up with the new term "purchasing agent mechanism".

3.2 External Statistics

Whether two Chinese characters can form a word, in addition to inferring the degree of combination between Chinese characters, the diversity of adjacent characters of a word is also a measure. Information entropy is usually used to measure the degree of freedom of the boundary. In information theory, entropy is used to express the mean value of the uncertainty of a random variable [23-26]. The

greater the entropy [27-28] of a random variable, the greater its uncertainty, and the more information it carries. Likewise, the less likely it is to estimate its value correctly.

The calculation formulas of left information entropy and right information entropy [29-32] are shown in (4) and (5)

$$H_l(W) = -\sum_{wl \in sl} p(w_l \mid w) \log_2 p(w_l \mid w). \quad \textbf{(4)}$$

$$H_r(W) = -\sum_{wr \in sr} p(w_r \mid w) \log_2 p(w_r \mid w). \quad \textbf{(5)}$$

Among them: s_l is the congregation of left adjacent words of candidate character w; s_r is the congregation of right adjacent words of candidate character w; $p(w_l|w)$ indicates the conditional probability that w_l is the left adjacent word of candidate word w; $p(w_r|w)$ indicates the conditional probability that w_r is the right adjacent word of the candidate character w.

The calculation formulas of $p(w_l|w)$ and $p(w_r|w)$ are show in (6)

$$p(w_r \mid w) = \frac{N(w, w_r)}{N(w)} \, p(w_l \mid w) = \frac{N(w_l, w)}{N(w)}. \quad \textbf{(6)}$$

Among them: $N(w_l|w)$ expresses the number of times w_l and w emerge together; $N(w)$ expresses the number of times w occurrences; This is the same principle that $N(w|w_r)$ expresses the number of times w_r and w emerge together.

Thus, if $H_l(w)$ is higher than the threshold, this demonstrates that the left boundary has been determined; The same that if $Hr(w)$ is higher than the threshold, this demonstrates that the right boundary has been determined.

3.3 Word Vector Representation

Natural language refers to the language normally used by humans. It can be understood naturally by humans, but it is difficult for computers to process. Generally, we need to convert natural language first and express it in the mathematical form [33]. Among them, word vector representation is a very good way of mathematically formalizing natural language symbols.

The word vector technology is based on a large amount of data to represent words as dense vectors, and this vector can indicate the degree of semantic similarity of words. For similar words, their corresponding word vectors are also similar [34]. The word2vec model is used for word vector training. The algorithm uses a large amount of text to create a high-dimensional word representation, captures the relationship between words, and does not require external annotations. After obtaining the segmented text of the corpus, it is converted into the training text format of word2vec, which is used as the text input for model training. The type of neural network used when training the model is the CBOW model, and the model structure is shown in Figure 1.

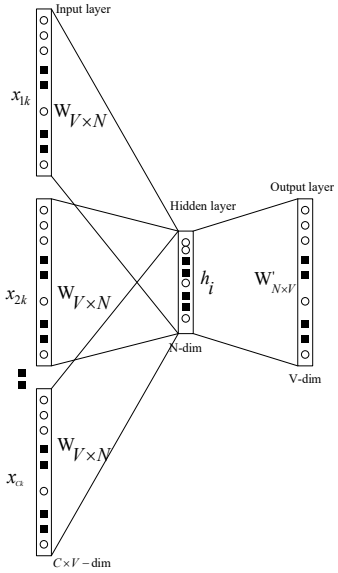

Figure 1. CBOW model

In the above structural diagram, the symbols x_{1k}, x_{2k}, …, x_{ck} represent the input words, which are actually a one-hot vector. By constructing a dictionary, building a word index can be easily achieved. The dimension of the vector is the same as the number of words in the dictionary. The value of only one dimension is 1, which corresponds to the position of the word in the dictionary index, and the value of the remaining dimensions is 0. The Hidden layer is obtained by looking up the table. First, initialize a words vector matrix W, where W is a two-dimensional matrix, the dimension of rows is equal to the number of words in the constructed dictionary, and depends on factors such as the size of the specific corpus.

The number of columns is a hyperparameters, which is artificially set. The specific network structure diagram is display in the following Figure 2:

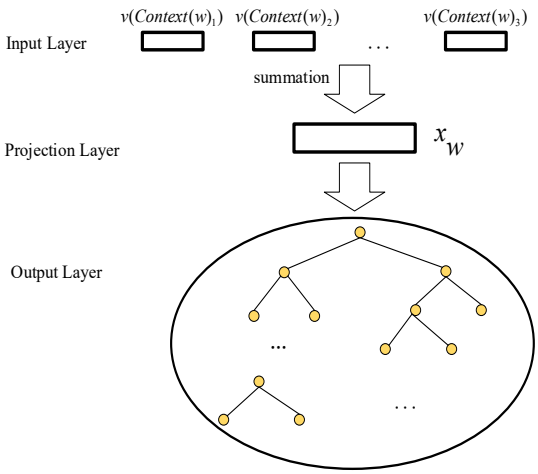

Figure 2. CBOW network structure

There are three layers of network structure included in the CBOW model: input layer, projection layer, and output layer. Take (*context*(*w*), *w*) as an example, Where *context*(*w*) is composed of *C* words adjacent to w. Input layer: the word vector $v(context(w)_1)$,$v(context(w)_2)$, ..., $v(context(w)_{2c})$ containing 2c words in context(w). Projection layer: the input layer 2c vectors are summed and accumulated $x_w = \sum_{i=1}^{2c} v(context(w)_i) \in R^m$, output layer: corresponding to a binary tree, which appears in the corpus words are used as leaf nodes, and the number of times each word appears in the corpus is used as the Huffman number constructed by the weight. There are a total of N=D dictionary size of leaf nodes, conform to each word in dictionary D, and there are N-1 non-leaf nodes.

Due to the particularity of the bid-winning contract in the tax-related field, there are a large number of synonyms in the corpus of candidate new words, and these synonyms are mistakenly separated when we do the word segmentation tool used in data preprocessing. Therefore, we need to find synonyms of candidate new words in the vectorized corpus according to the word vector similarity principle and expand the new word dictionary.

Cosine distance, also known as Cosine Similarity, is used to measure the cosine of the angle between two vectors in a multi-dimensional space. In a right-angle co-ordinate system, assuming that the *a* a vector is (x_1, y_1) and the *b* vector is (x_2, y_2), then the cosine vector can be expressed in the form of Figure 3. A high degree of similarity is indicated when the angle between the vectors *a* and *b* is smaller, and when the angle is 0, i.e. the two vectors are equal. In contrast to the Euclidean distance, the cosine distance is more concerned with the difference between the two vectors at the directional level.

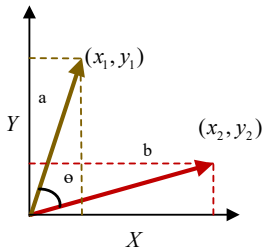

Figure 3. Cosine theorem space representation

For two multidimensional vectors $A = (x_1, x_2, x_3, ..., x_n)$ and $B = (y_1, y_2, y_3, ..., y_n)$, the formula for the cosine similarity of the angle between them is shown in (7).

$$sim = \cos\theta = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sum_{i=1}^{n} y_i^2}. \quad (7)$$

where x_i and y_i are the corresponding components of the multidimensional vectors *A* and *B* in space, respectively.

4 Our Proposed Word Discovery Scheme Based on Word Vector Representation

4.1 Algorithm Flow

This study is a new word detection algorithm for bidding contracts in tax-related fields. Because the bidding contract corpus contains a massive sparse data, First, the corpus is preprocessed, and then the preprocessed data is segmented by using the Jieba word segmentation system to obtain the initial scattered string, measure the degree of combination of words within the scattered string set, expand and filter the word strings larger than the threshold, then, use external statistics to determine the final result of the new word and finally use the cosine similarity principle in the word vector to enlarge the new word dictionary.

The algorithm flow is shown in Figure 4.

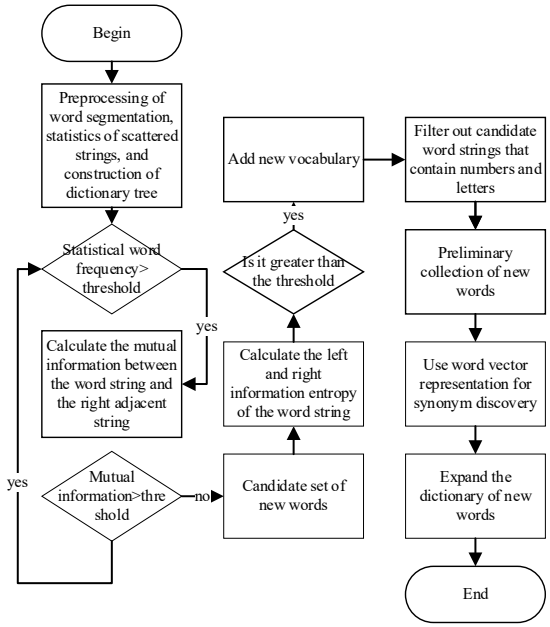

Figure 4. Algorithm flow chart

5 Performance Analysis

The corpus used in this experiment is the one hundred thousand bid-winning webpages published on the Chinese Government Procurement Network. The algorithm evaluation indicators used in this article are accuracy P (precision), recall rate R (recall), and F-measure (F-measure).

$$p = \frac{N \cap M}{N} \times 100\%. \quad (8)$$

$$R = \frac{N \cap M}{M} \times 100\%. \quad (9)$$

$$F = \frac{2PR}{P+R}. \quad (10)$$

Among them: N indicates the number of new words correctly identified; M indicates the total number of word strings in this experiment.

In order to verify the effectiveness of the algorithm in this article, two comparative experiments have been added: one is based on the traditional word segmentation tool jieba to segment the corpus, and the algorithm is based on the combination of the N-gram algorithm and word-internal combination degree and boundary freedom and the algorithm of this paper ME-Vector Compare. The experimental results are shown in Table 1. The histograms in Figure 5 and Figure 6 show the experimental results more intuitively.

It can be seen from Table 1 and Figure 5 and Figure 6 that the new word detection method of word vector features that integrates internal and external statistics proposed in this research has reached excellent consequences in the process of compound word discovery in the tax-related financial field. Its accuracy, recall, and F The value is improved compared to other algorithms. The first comparison experiment is to segment the corpus based on the traditional jieba word segmentation algorithm. The jieba word segmentation algorithm uses the built-in dic.txt dictionary to generate a trie tree, and then treats the segmented sentence, and generates a trie tree based on the dict.txt Directional acyclic graph (DAG), on the basis of this word graph, uses dynamic programming algorithm to generate the best path for segmentation, uses HMM model to identify unregistered words, and finally recalculates the segmentation path. Due to the particularity of the tax-related financial field, many new words are combined words. Therefore, it is difficult to identify combined new words using the jieba word segmentation algorithm without a domain dictionary. For example, the combined new word “winning bid amount” is using the jieba algorithm Time was mistakenly divided into two random strings of “winning bid” and “amount”; therefore, its accuracy rate is relatively low compared to the other two experiments.

The second comparative experiment is to use the N-gram algorithm to combine the scattered strings based on the segmentation of the word segmentation tool, and then obtain the candidate new word set by analyzing the internal combination of words and boundary freedom between adjacent scattered strings. In the case that some new words are misclassified by the word segmentation tool, the N-gram algorithm is used to combine the scattered string sets after the word segmentation, which solves the problem of multi-character recognition, such as the combination of new words “winning bid amount” After being mistakenly divided into “winning bid” and “amount” by the word segmentation tool, the candidate string of “winning bid amount” is combined under the N-gram algorithm, and then the internal word-formation probability and external word-formation probability of the word are considered at the same time Next, the combined new word “winning bid amount” was identified. Although the number of correct new words that can be correctly identified using the N-gram+MI+BE method increases, it is inevitable that there will be many rubbish word strings in the new word set, resulting in low accuracy.

Table 1. New word discovery experiment results

Methods	Precision/ %	Recall/ %	F-measure
Jieba	69.17	82.65	75.31
N-gram+MI+BE	72.82	84.47	78.21
ME-Vector	74.68	85.49	79.72

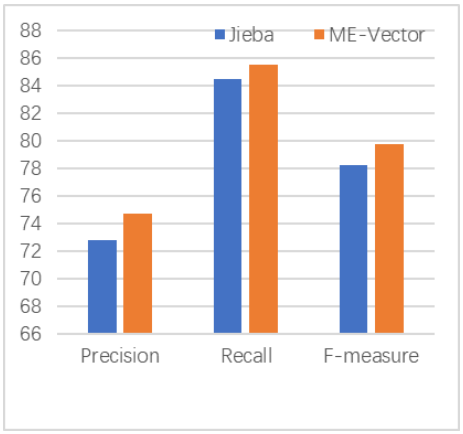

Figure 5. Comparative experiment 1

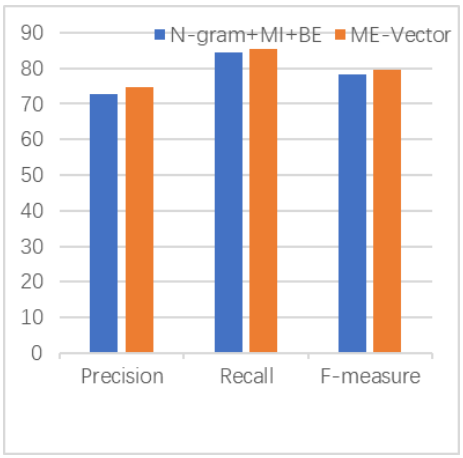

Figure 6. Comparative experiment 2

In the process of identifying new words and their compound words in tax-related fields, because there are many similar words in the corpus, the N-gram+MI+BE method cannot find the low-frequency words in the corpus well. For the subsequent text. The information extraction process has caused great problems, so the method in this paper uses the CBOW model in word2vec to train and represent the word vector, and then measures the similarity between the two word embedding vectors according to the cosine similarity, and then expands the domain dictionary. It can be seen from the results in Figure 7 that when the context sliding window is set to 3, the number of new words similar to valid words is overall higher than that when the window is 5, and as the word vector dimension increases from 25 to 200 and then to 300, The number of similar words in new words

generally showed a trend of rising first, then maintaining a steady or even falling trend. And it reaches the peak when the dimension is 200, but in conjunction with the similarity distribution in Figure 8, it can be seen that when the sliding window is fixed at 3 above when the dimension is 200, the total number of effective words is the highest, but the number of similarities is far below 50%. It is much larger than the case where the dimension is 100, and the similarity distribution of more than 50% is also far lower than the case where the dimension is 100. This is mainly because the dimension of the word vector represents the characteristics of the word. The increase in the dimension of the word vector in the early stage can be more Richly express the semantic information of words, to better realize the distinction of words, and the subsequent word vector scoring can also better count the collection of closely related words. However, if the dimensionality of the word vector is too high, the relationship between words will be too weakened, and the measure of similarity between words will be reduced. In addition, a too high word vector dimension can also lead to overfitting. Therefore, this article sets the dimension of the word vector to 100 and the sliding window of the context to 3.

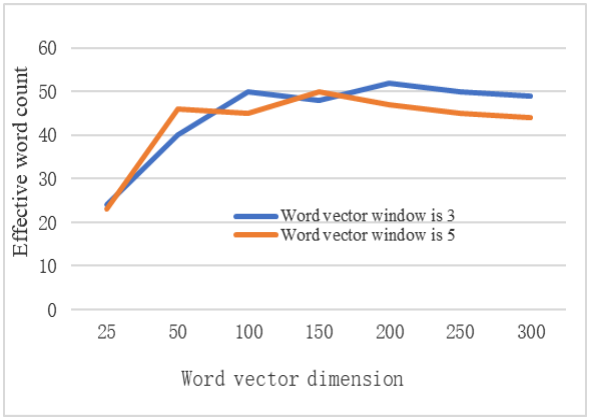

Figure 7. The number of effective words in different word vector dimensions and context windows

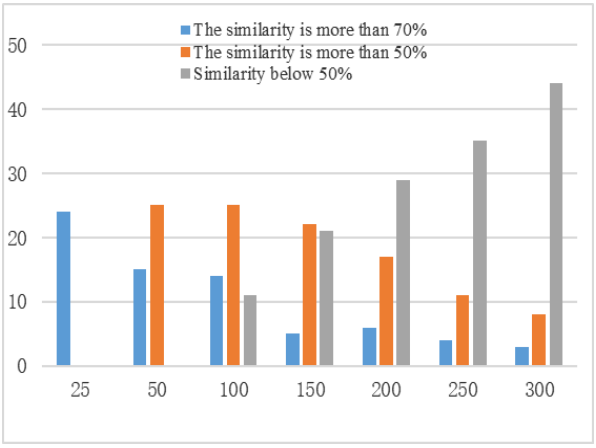

Figure 8. Similarity under different word vector dimensions

When the cosine distance is used to calculate the similarity between words, many new words will be recognized. For example, the synonyms of "winning bid amount", "winning bid price", "bid price", and "winning bid price" cannot be recognized at all in the N-gram+MI+BE algorithm, and in the corpus represented by word vector, according to the similarity The principle can solve this problem well, and some long words "name of purchasing agency" and "name of centralized purchasing agency" are also identified in the similar words of "name of agency".

When using the similarity metric to mine low-frequency synonymous neologisms, many new words are identified. The results of the similarity word identification component of the winning supplier are shown in Table 2.

Table 2. Winning bid supplier synonym display

Original word string	Synonyms
Winning supplier	Sold unit, vendor candidate, vendor, successful candidate, first sold candidate, name of sold unit, successful bidder, name of successful bidder Name of the successful bidder, name of the transaction unit, supplier candidate, winning bidder

6 Conclusion

In the process of natural language processing, word segmentation is the bottom work. The accuracy of word segmentation is largely constrained by some unrecognized new words, which also has a huge impact on the subsequent text extraction and entity recognition. Considering that there are sparse data and a large number of combined new words in the bidding contract of tax-related field, this paper uses the statistical method of internal combination degree and boundary freedom degree to find new words in the financial tax related field based on Jieba word segmentation tool and finally expands the domain dictionary by cosine similarity measurement in the vectorized corpus formed by n-gram. In this paper, through experiments, using a me-vector algorithm, through reasonable adjustment of parameters, we can also achieve good results on the small-scale corpus. This paper is a preliminary work in this field. In the future, we will continue to increase the corpus of training word vectors, and combine deep learning and named entity recognition to further improve the accuracy of long entity recognition.

Acknowledgement

This job is supported by National Key R&D Program of China (No. 2022YFE0138600). This job is also supported by Supported by Natural Science Foundation of Shaanxi Province of China (2021JM-344) and Open Fund for Chongqing Key Laboratory of Computational Intelligence (No. 2020FF02) and Shaanxi Key Laboratory of Intelligent Processing for Big Energy Data (No. IPBED7).

References

[1] H. Liu, P. D. Gao, Y. Xiao, New words discovery method based on word segmentation result, *2018 IEEE ACIS 17th International Conference on Computer and Information Science (ICIS)*, Singapore, 2018, pp. 645-648.

[2] W.-F. Wang, H. Xu, W. Yang, X. Wu, Review of Chinese word segmentation algorithms, *Group Technology & Production Modernization*, Vol. 35, No. 3, pp. 1-8, September, 2018.

[3] D. Kerremans, J. Prokić, Mining the web for new words: Semi-automatic neologism identification with the NeoCrawler, *Anglia*, Vol. 136, No. 2, pp. 239-268, June, 2018.

[4] Y. Qian, Y. Du, X. Deng, B. Ma, Q. Ye, H. Yuan, Detecting new Chinese words from massive domain texts with word embedding, *Journal of Information Science*, Vol. 45, No. 2, pp. 196-211, April, 2019.

[5] H. Q. Gheni, A. M. Hussein, W. K. Oleiwi, Suggesting new words to extract keywords from title and abstract, *International Journal of Electrical and Computer Engineering*, Vol. 9, No. 5, pp. 4441-4445, October, 2019.

[6] X. Chen, C. Han, Y. An, L. Liu, Z. Li, R. Yang, Extracting New Words with Mutual Information and Logistic Regression, *Data Analysis and Knowledge Discovery*, Vol. 3, No. 8, pp, 105-113, August, 2019.

[7] Y. Luo, Y. Zuo, J. Wu, H. Li, Online Detection of Domain-Specific New Words in Text Streams, *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*, 2018, Hangzhou, China, pp. 1-6.

[8] D. Klop, L. Marais, A. Msindwana, F. D. Wet, Learning new words from an interactive electronic storybook intervention, *The South African Journal of Communication Disorders*, Vol. 65, No. 1, pp. 1-8, September, 2018.

[9] X.-Z. Fu, S.-Q. Liu, Y. Liu, Z.-W. Guo, M.-L. Zhao, Research on the Method of Stable New Words Identification Based on the Law of Survival, *Journal of Xinjiang University (Natural Science Edition)*, Vol. 35, No. 1, pp. 73-79, February, 2018.

[10] W. Li, K. Guo, Y. Shi, L. Zhu, Y. Zheng, DWWP: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain, *Knowledge-Based Systems*, Vol. 146, pp. 203-214, April, 2018.

[11] J. Liu, F. Wu, C. Wu, Y. Huang, X. Xie, Neural Chinese word segmentation with dictionary, *Neurocomputing*, Vol. 338, pp. 46-54, April, 2019.

[12] Y. Liu, M. Zhang, Neural network methods for natural language processing by Yoav Goldberg, *Computational Linguistics*, Vol. 44, No. 1, pp. 193-195, April, 2018.

[13] F. Chen, Y.-Q. Liu, C. Wei, Y.-L. Zhang, M. Zhang, S.-P. Ma, Open Domain New Word Detection Using Condition Random Field Method, *Journal of Software*, Vol. 24, No. 5, pp. 1051-1060, May, 2013.

[14] C. Zhang, S. Zhao, Y. He, An integrated method of the future capacity and RUL prediction for lithium-ion battery pack, *IEEE Transactions on Vehicular Technology*, Vol. 71, No. 3, pp. 2601-2613, March 2022.

[15] Y.-F. Huang, M. Liu, X.-Y. Li, New Words Detection Research for Programming Domain, *Proceedings - 2021 6th International Symposium on Computer and Information Processing Technology*, Changsha, China, 2021, pp. 777-782.

[16] R.-P. Yao, G.-Y. Xu, J. Song, Micro-blog new word discovery method based on improved mutual information and branch entropy, *Journal of Computer Applications*, Vol. 36, No. 10, pp. 2772-2776, October, 2016.

[17] K. Zhang, Q. Liu, H. Zhang, X.-Q. Cheng, Automatic recognition of Chinese unknown words based on roles tagging, *Proc. of the 1st SIGHAN Workshop on Chinese Language Processing*, Taipei, Taiwan, 2002, pp. 71-78.

[18] A. Davy, T. Ehret, J. M. Morel, P. Arias, G. Facciolo, Video denoising by combining patch search and CNNs, *Journal of Mathematical Imaging and Vision*, Vol. 63, No. 1, pp. 73-88, January, 2021.

[19] W. Wei, X. Xia, W. Marcin, X. Fan, D. Robertas, Y. Li, Multi-sink distributed power control algorithm for Cyber-physical-systems in coal mine tunnels, *Computer Networks*, Vol. 161, pp. 210-219, October, 2019.

[20] W. Wei, H. Song, W. Li, P. Shen, A. Vasilakos, Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network, *Information Sciences*, Vol. 408, pp. 100-114, October, 2017.

[21] P. Naveen, P. Sivakumar, Adaptive morphological and bilateral filtering with ensemble convolutional neural network for pose-invariant face recognition, *Journal of Ambient Intelligence and Humanized Computing*, Vol. 12, No. 11, pp. 10023-10033, November, 2021.

[22] S. Liu, X. Xu, Y. Zhang, K. Muhammad, W. Fu, A Reliable Sample Selection Strategy for Weakly Supervised Visual Tracking, *IEEE Transactions on Reliability*, Vol. 72, No. 1, pp. 15-26, March, 2023.

[23] S. Zhao, C. Zhang, Y. Wang, Lithium-ion battery capacity and remaining useful life prediction using board learning system and long short-term memory neural network, *Journal of Energy Storage*, Vol. 52(B), Article No. 104901, August, 2022.

[24] L. Li, H. Li, G. Kou, D. Yang, W. Hu, J. Peng, S. Li, Dynamic camouflage characteristics of a thermal infrared film inspired by honeycomb structure, *Journal of Bionic Engineering*, Vol. 19, No. 2, pp. 458-470, March, 2022.

[25] Y. Y. Ghadi, I. Akhter, H. Aljuaid, M. Gochoo, S. A. Alsuhibany, A. Jalal, J. Park, Extrinsic behavior prediction of pedestrians via maximum entropy Markov model and graph-based features mining, *Applied Sciences*, Vol. 12, No. 12, Article No. 5985, June, 2022.

[26] W. Wei, X. Fan, H.-B. Song, X.-F. Fan, J.-C. Yang, Imperfect information dynamic stackelberg game based resource allocation using hidden Markov for cloud computing, *IEEE Transactions on Services Computing*, Vol. 11, No. 1, pp. 78-89, January-February, 2018.

[27] C. Zhang, S. Zhao, Z. Yang, Y. Chen, A reliable data-driven state-of-health estimation model for lithium-ion batteries in electric vehicles, *Frontiers in Energy Research*, Vol. 10, Article No. 1013800, September, 2022.
[28] G. U. Bhargava, S. V. Gangadharan, FPGA implementation of modified recursive box filter-based fast bilateral filter for image denoising, *Circuits, Systems, and Signal Processing*, Vol. 40, No. 3, pp. 1438-1457, March, 2021.
[29] B. Goyal, A. Gupta, A. Dogra, D. Koundal, An adaptive bitonic filtering based edge fusion algorithm for Gaussian denoising, International Journal of Cognitive Computing in Engineering, Vol. 3, pp. 90-97, June, 2022.
[30] W. Wei, Y. Qi, Information potential fields navigation in wireless Ad-Hoc sensor networks, *Sensors*, Vol. 11, No. 5, pp. 4794-4807, May, 2011.
[31] M. S. Mahdi, A. J. Mohammed, M. M. Jafer, Unusual activity detection in surveillance video scene: Review, *Journal of Al-Qadisiyah for Computer Science and Mathematics*, Vol. 13, No. 3, pp. 92-98, September, 2021.
[32] W. Wei, Q. Ke, J. Nowak, M. Korytkowski, R. Scherer, M. Woźniak, Accurate and fast URL phishing detector: A convolutional neural network approach, *Computer Networks*, Vol. 178, Article No. 107275, September, 2020.
[33] M. Wu, L. Tan, N. Xiong, A structure fidelity approach for big data collection in wireless sensor networks, *Sensors*, Vol. 15, No. 1, pp. 248-273, January, 2015.
[34] H. Li, J. Liu, K. Wu, Z. Yang, R. W. Liu, N. Xiong, Spatio-temporal vessel trajectory clustering based on data mapping and density, *IEEE Access*, Vol. 6, pp. 58939-58954, August, 2018.

Biographies

Wei Wei received his Ph.D. and M.S. degrees from Xi'an Jiaotong University in 2011 and 2005, respectively. Currently he is an associate Professor at Xi'an University of Technology. He currently is an IEEE Senior Member. His research interests include Wireless Networks and Wireless Sensor Net-works Application, Mobile Computing, Distributed Computing, and Pervasive Computing.

Wei Liu graduated from Shaanxi University of Science and Technology with a bachelor's degree, majoring in information and computing science. She studied in Xi'an University of Science and Technology with a master's degree in natural language processing.

Beibei Zhang received the MS and PhD degrees from the Xian Jiaotong University, respectively. He is currently an assistant professor at Xian University of Technology. His research interests include wireless networks and wireless sensor networks applications, mobile computing, distributed computing, and pervasive computing.

Rafał Scherer received his MSc degree in computer science from the Czestochowa University of Technology, Poland, in 1997 and his PhD in 2002 from the same university. Currently, he is an associate professor at Czestochowa University of Technology. His present research interests include machine learning and neural networks for image processing, computer system security, prediction and classification.

Robertas Damaševičius graduated at the Faculty of Informatics, Kaunas University of Technology (KTU) in Kaunas, Lithuania in 1999, where he received a B.Sc. degree in Informatics. He finished his M.Sc. studies in 2001 (cum laude), and he defended his Ph.D. thesis at the same University in 2005. Currently, he is a Professor at Department of Software Engineering, KTU and lectures robot programming and software maintenance courses.