Enhanced New Words Recognition Based on Multi-level Semantic Vectors and Multi-task Learning Models

Jin Pan[1], Yang Chen[1], Chunlu Zhao[1], Yang Liu[1], Jie Chu[2]*

[1] National Computer Network Emergency Response Technical Team/Coordination Center of China, China
[2] Institute of Network Technology (Yantai), China
jinpancert@163.com, cylovehehe@163.com, chunluzhao@cert.org.cn,
liuyangcert@163.com, chujie@int-yt.com

Abstract

Neologism discovery is a basic task in natural language processing, and it is very important to improve the performance of various downstream tasks. In order to solve problems such as word segmentation errors easily caused by existing technologies and incomplete capture of word semantic information, this paper proposes an enhanced new words recognition based on multi-level semantic vectors and multi-task learning models, aiming to solve the problems of word segmentation errors, incomplete semantic capture and dynamic recognition in existing technologies. First, an improved hash algorithm is used to generate dictionaries. A multi-level time series model is used to identify potential neologism candidates and map them to a high-dimensional vector space to generate synthetic semantic vectors. Then, a context-semantic graph model is constructed to analyze the context compatibility of words, and the sentiment score and domain relevance score are calculated through the multi-task learning model. Finally, the comprehensive score is used to identify new words. Experimental results show that this method has significant advantages in accuracy, semantic understanding and application range.

Keywords: New word discovery, Word segmentation, Multi-tasking learning, Multi-level semantic vector

1 Introduction

With the popularization and development of the Internet, social media has become an important channel for information dissemination, which greatly affects people's daily lives [1-2]. In this process, a large number of Internet neologisms have been quickly spread and widely used with the help of social media. These new words are not only the induction and summary of social hot phenomena, but also the similar substitution of specific words. Although this linguistic phenomenon enriches text expression, it also brings new challenges to natural language processing tasks. Words are contained in the text semantic information and are able to use the smallest structural unit of the independent, so word segmentation (CWS) is a natural language processing (NLP) foundation, and its performance will have a direct impact on the effect of NLP downstream tasks. Early word segmentation methods mainly include mechanical word segmentation [3] and statistical word segmentation [4]. Mechanical word segmentation requires the construction of a large enough word list in advance. Segmentation of sentences occurs by setting the combination rules of words in the word list. Statistical word segmentation is based on the co-occurrence frequency of adjacent words to calculate the confidence of their formation of words, without the need to pre-construct the word list. Because the word segmentation models used by these two methods are relatively simple, they cannot describe the complex word formation law well, so the part-of-speech segmentation is not ideal.

In recent years, with the continuous development of NLP technology, especially the introduction of deep learning methods, there have been many innovations in the field of word segmentation. By building a multi-layer neural network model, deep learning can automatically learn complex language features from data, which greatly improves the accuracy and flexibility of word segmentation. These methods can be more flexible and accurate by automatically learning semantic information and context in a large amount of text data. For example, models based on structures such as convolutional neural networks (CNN) [5], recurrent neural networks (RNN) [6] and Transformers [7] can perform word segmentation in more complex linguistic environments and show greater adaptability in dealing with polysemy and ambiguity resolution. Especially in terms of finding new words and filtering noisy word strings, the neural network model can capture potential word boundaries through dynamic learning of context information, and improve the adaptability to language diversity. However, while deep learning methods have made significant progress in many application scenarios, there are still problems such as how to use the information in the existing corpus more efficiently, deal with uncertainties, and optimize model complexity.

2 Related Works

Current neologism discovery methods can be divided

*Corresponding Author: Yang Chen; E-mail: cylovehehe@163.com
DOI: https://doi.org/10.70003/160792642025052603005

into two categories: unsupervised and supervised. Unsupervised neologism discovery methods usually require the support of large-scale corpus, and do not require the construction of annotated data sets compared with supervised methods. Unsupervised methods fall into two categories: rule-based and statistics-based. The rule-based neologism discovery method refers to the use of language word formation rules and contextual relationships to discover new words, the core of which is to accurately discover word formation rules and then match word sequences. The accuracy of this method is high, but the coverage of the rules is limited. It has poor applicability and low portability. The key point of rule-based approach is to mine the word formation features, parts of speech features or semantic information of new words, so as to establish a rule base, a pattern base or a professional thesaurus, and then give potential new words in the way of rule matching. Zhao et al. [8] built a domain syntax dictionary based on dependency syntax analysis and TF_IDF, and word vector calculated the similarity between candidate new words and the entered words in the dictionary, so as to complete the judgment of domain new words. The advantage of the rule-based method is accurate identification, but the disadvantage is difficult rule summary and poor migration ability.

The new word discovery method based on statistics is to find new words in large-scale experimental corpus by statistical model. Statistical machine learning methods can do a good job of labeling and classifying data. Feng et al. [9] proposed an automated new word discovery technique based on four statistical indicators: word frequency, mutual information, left and right information entropy, and inverse document frequency. Li [10] combined the improved multi-PMI algorithm with the double threshold word segmentation method, optimized the N-gram model with branch entropy, and improved the accuracy rate, recall rate and F-value of new word recognition. Wei et al. [11] proposed an unsupervised neologism discovery method combining statistical features and word vector representation to improve the effectiveness of neologism recognition in tax-related fields by automatically expanding dictionaries. Duan et al. [12] proposed a vocabulary construction method for professional domains based on new word recommendation to solve the sparsity problem of unsupervised neologism discovery algorithm by enhancing mutual information, branching entropy and NC value. In order to make full use of context features, Tian et al. [13] proposed the neural network framework WM-SEG, which uses memory networks to integrate word formation information into word segmentation networks. Chinese word segmentation models tend to rely on word lists to learn word segmentation knowledge. Lin et al. [14] proposed a context-aware approach that combines unsupervised sentence representation learning with a multi-criterion training framework. Aiming at the problem of multi-standard word segmentation, Qiu et al. [15] proposed a unified multi-label model based on Transformer encoder and multiple standard word segmentation tags, which further improves the party segmentation capability under each standard. In order to improve the accuracy of named entity recognition in the medical field, Liu et al. [16] proposed the PEM model, which combines BERT [17], graph attention network and multi-head cross-attention mechanism to fuse semantic and adjacent-dependent features. Li et al. [18] proposed the CWSeg method, which enhanced the Chinese word segmentation system based on pre-trained language models by developing cooperative training and multifunctional decoding strategies.

The existing technology has many limitations in the task of new word recognition and word segmentation. First of all, due to the constant update and change of new words on the Internet, they are often not fully covered by traditional segmentation models and dictionaries, resulting in segmentation errors. In addition, the existing methods fail to fully capture the frequency difference between words in the short and long term when dealing with new words on the Internet, which further leads to the low accuracy of abnormally high frequency words recognition. Because of the failure to dynamically capture the expression of words in different contexts, the effect of polysemy disambiguation is not satisfactory. These problems ultimately affect the reliability and accuracy of new word recognition.

3 The Proposed Approach

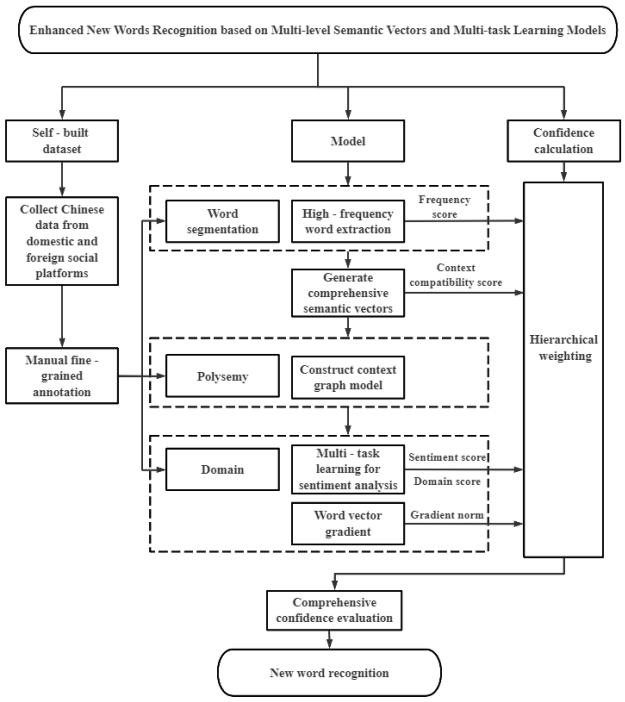

Figure 1. The framework of the proposed approach

In order to solve the problems of word segmentation errors, poor context adaptability and incomplete semantic capture in prior art, this paper proposes a new enhanced new words recognition based on multi-level semantic vectors and multi-task learning models. By improving the hash algorithm and multi-level time series model, the segmentation accuracy is improved, and the context information of words is captured by generating comprehensive semantic vectors. The improved context

semantic graph and self-supervised learning graph convolutional network are used to further improve the accurate recognition of new words, especially in polysemy disambiguation and context compatibility analysis. Through the multi-task learning model, combined with emotion analysis and correlation judgment of specific fields, the text is deeply understood, and the accuracy of neologism judgment is improved. Finally, combined with the multi-level weighted confidence evaluation mechanism, this method effectively improves the reliability and accuracy of new word recognition. The framework of the proposed approach is shown in Figure 1.

3.1 High-frequency Word Extraction Based on Improved Hashing

The section first acquires text data and performs preprocessing and tokenization. An improved hashing algorithm is used to calculate the hash value for each word, forming a dictionary. A multi-level time series model is then employed to compute short-term and long-term frequencies, identify abnormally high-frequency words as potential new word candidates, and map them into a high-dimensional vector space to generate comprehensive semantic vectors.

To acquire text data, a distributed computing framework is used for preprocessing the input text data. After preprocessing, tokenization is performed, splitting the text data into multiple parts that are processed in parallel on different computing nodes. The input text is segmented into words or phrases, with stop words removed. The hash value for each word or phrase is calculated, and words or phrases with the same hash value are grouped together to form a dictionary. This results in a dictionary containing all words and their corresponding hash values.

Specifically, an improved hash algorithm is used to segment text data. The improved hash algorithm improves the accuracy of word segmentation by increasing the weight and position parameters of words. The specific formula is as follows:

$$h(w_i) = \sum_{s=1}^{S} c_s \cdot p_s(w_i)) \bmod m \tag{1}$$

Where $h(w_i)$ represents the hash value of the i-th word w, S is the length of the word, c_s is the weight of the s-th position, $p_s(w_i)$ is the hash value of the character of the i-th word in the s-th position, and is the size of the hash table.

After getting the dictionary, count the frequency of each word. In order to capture the changes of words in time, a multi-level time series model is constructed to identify abnormally high frequency words through short- and long-term word frequency changes. Specifically, the short-term frequency of $F_s(w_i, t)$ and long-term frequency of $F_l(w_i, t)$ are calculated as follows:

$$F_s(w_i,t) = \frac{1}{N_s(t)} \sum_{k=1}^{N_s(t)} \delta(w_i, w_k) \cdot \cos(\omega_s t + \phi_s) \tag{2}$$

$$F_l(w_i,t) = \frac{1}{N_l(t)} \sum_{k=1}^{N_l(t)} \delta(w_i, w_k) \cdot \cos(\omega_l t + \phi_l) \tag{3}$$

Where, $F_s(w_i, t)$ and $F_l(w_i, t)$ represent the frequency of the word w_i in short and long time t respectively, $N_s(t)$ and $N_l(t)$ are the total number of texts in short and long term, $\delta(w_i, w_k)$ is the indicator function, 1 when $w_i = w_k$, 0 otherwise, ω_s and ω_l are the frequency, ϕ_s and ϕ_l are the phase offset.

Calculate the difference between the frequencies of words in the short and long term to identify changes in abnormally high frequencies. Define the frequency difference measure $\Delta F(w_i, t)$ as follows:

$$\Delta F(w_i,t) = \left| F_s(w_i,t) - F_l(w_i,t) \right| \tag{4}$$

Frequency differences are standardized for better comparison. z-score can be used to standardize:

$$z(w_i,t) = \frac{\Delta F(w_i,t) - \mu_{\Delta F}}{\sigma_{\Delta F}} \tag{5}$$

Where, $z(w_i, t)$ is the standardized frequency difference, and $\mu_{\Delta F}$ and $\sigma_{\Delta F}$ respectively represent the mean and standard deviation of the frequency difference of all words at time t.

A threshold T_z is set. If the standardized frequency difference $z(w_i, t)$ exceeds the threshold T_z, it is considered that the current word has abnormal high frequency phenomenon in time t. All words w_i meeting the conditions are output as abnormal high frequency words, so as to identify words with abnormal high frequency in a specific period of time as potential new word candidates.

3.2 High-frequency Word Meaning Information Extraction

In order to capture the semantic information of words more comprehensively, this paper combines the context information and paragraph information, and maps the abnormal high-frequency words and their context information and paragraph into the high-dimensional vector space. The context information refers to a certain number of words adjacent to the abnormally high frequency words (that is, the target word w_i) in the sentence. Set the context window size to c, then the context word set $\{u_1, u_2, \ldots, u_c\}$ contains c words before and after w_i. The paragraph refers to the semantic vector of the paragraph where the target word w_i is located. Assume that the vector of the paragraph in which the target word is located is d_i.

The specific implementation is to map the context information and paragraph into the low-dimensional vector space through word2vec and sen2vec, and obtain the semantic vector representing $\{\vec{u}_j, \vec{d}_i\}$, where $\vec{u}_j$ is the embedding vector of the j-th context word, and $\vec{d}_i$ is the embedding vector of the paragraph where the target word w_i is. Through a multi-layer perceptron, the low-

dimensional vector is further processed to generate the high-dimensional synthetic semantic vector. The specific formula is as follows:

$$\vec{v}_{w_i} = \sum_{j=1}^{c} \alpha_{1j} \cdot \tanh\left(\beta_{1j} \cdot \vec{u}_j + \gamma_{1j}\right) + \alpha_2 \cdot relu\left(\beta_2 \cdot \vec{d}_i + \gamma_2\right) \quad (6)$$

Where $\vec{v}_{w_i}$ represents the synthetic vector representation of the word w_i, α_{1j} represents the weight of the embedding vector of the context word in the synthetic semantic vector, β_{1j} and γ_{1j} are respectively the linear transformation coefficient and bias of the embedding vector $\vec{u}_j$ of the context word before the activation function through tanh, α_2 represents the weight of the paragraph embedding vector $\vec{d}_i$ in the synthetic semantic vector. β_2 and γ_2 represent the linear transformation coefficient and bias of the paragraph embedding vector before $\vec{d}_i$ is activated by the function, respectively.

After the synthetic semantic vectors are generated, a clustering algorithm is used to cluster the semantic vectors in the high-dimensional space to identify potential neologism clusters. The multi-view clustering algorithm calculates the similarity between each vector through multiple iterations. The cosine similarity algorithm is used to calculate the similarity. The algorithm formula is as follows:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \quad (7)$$

Where A and B are two vectors, ||A|| and ||B|| represent the modulo (length) of A and B, respectively, and A·B represents the dot product of A and B. Finally, the cluster center is formed, and the attribution of each vector is updated according to the change of the cluster center. The specific steps of the clustering algorithm are as follows:

(1) Initial cluster centers: randomly select a preset number of initial cluster centers.

(2) Calculating similarity: use cosine similarity to calculate the similarity between each vector and the cluster center.

(3) Update clustering: each vector is assigned to the most similar cluster center according to its similarity.

(4) Update center: recalculate the center vector of each cluster.

(5) Iteration: repeat the similarity calculation, update the cluster, and update the center until the clustering results converge.

3.3 Disambiguation Algorithm Based on Context Semantic Graph Model

In this section, comprehensive semantic vectors are used to build a contextual semantic graph model, analyze the context compatibility of words, build a multi-task learning emotion analysis model, and calculate emotion scores and domain-specific relevance scores respectively. Through comprehensive calculation of each score, the total confidence of each candidate word is obtained, and potential new words are identified.

After the potential neologism clusters are identified, the words in the potential neologism clusters are used as nodes to construct the improved contextual semantic graph model. Specifically, each word and its context information in a potential neologism cluster will be used to generate nodes and edges of a semantic graph, build an improved contextual semantic graph model, and analyze the compatibility of words in multiple contexts. The contextual semantic graph model takes into account the semantic relationships of sentences, paragraphs and full text, and can analyze the expression of words in different contexts more comprehensively. A self-supervised learning graph convolutional network is used to embed nodes in semantic graphs and disambiguate polysemous words. The specific formula is as follows:

$$S_{cp}(w_i) = \frac{1}{|N(w_i)|} \sum_{g \in M(w_i)} \frac{\vec{v}_{w_i} \cdot \vec{v}_g}{\|\vec{v}_{w_i}\| \cdot \|\vec{v}_g\|} + \lambda \cdot sigmoid(\vec{v}_{w_i} \cdot \vec{v}_g) \quad (8)$$

Where, $S_{cp}(w_i)$ represents the context compatibility score of the word w_i , $M(w_i)$ is the set of neighbor nodes of the word w_i, $\vec{v}_g$ is the synthetic vector representation of the neighbor node g of the word w_i, and λ is the adjustment parameter. Using the above formula, it is possible to calculate the compatibility score of each word in its context and identify potential new words that are incompatible with the context. The specific construction steps of the improved contextual semantic graph model are as follows: each word and its context in a potential neologism cluster are represented as nodes in the context semantic graph, and edges between nodes represent semantic relationships between words. Each node is initialized to its synthetic semantic vector. The graph is embedded with a graph convolutional network and the vector representation of each node is updated. The context compatibility score for each word is calculated based on the updated node vector.

3.4 Multitask Classification Recognition and Decision Method

This section uses ERNIE3.0 to construct a multi-task learning emotion and domain analysis model to classify emotions and determine domain-specific relevance. ERNIE3.0 integrates autoregressive network and self-coding network, and because of the introduction of large-scale knowledge graph data, the model can achieve excellent performance in understanding task, generation task, zero-sample learning task and common-sense reasoning task. The basic framework is shown in the figure below. The core features are multi-paradigm unified training and general knowledge text prediction, and a large amount of existing business text data is used to fine-tune the model. The ERNIE3.0 basic framework is shown in Figure 2.

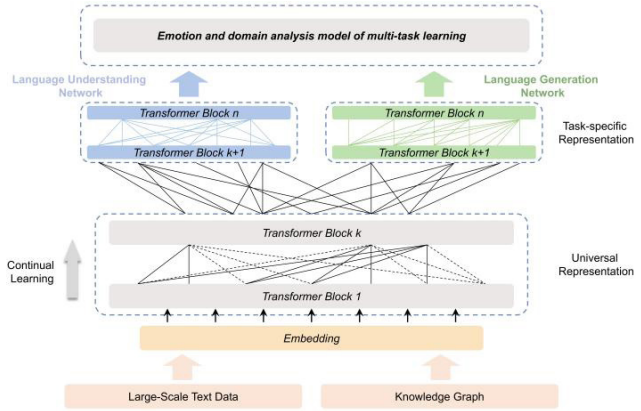

Figure 2. ERNIE3.0 basic framework

The multi-task learning model integrates the synthetic vector representation information, which can analyze the emotional tendency of words and their contexts more accurately. The cross-entropy loss function is used to optimize the sentiment analysis model to calculate the sentiment scores of words and their contexts and the results of domain-specific relevance judgment. The specific formula is as follows:

$$S_{em}(w_i) = \sum_{p=1}^{P} \lambda_p \cdot \sigma(\vec{\alpha}_p \cdot \vec{v}_{w_i} + \theta_p) \tag{9}$$

$$S_{ha}(w_i) = \sum_{q=1}^{Q} \lambda_q \cdot soft\max(\vec{\alpha}_q \cdot \vec{v}_{w_i} + \theta_q) \tag{10}$$

Where, $S_{em}(w_i)$ represents the emotion score of the word w_i, $S_{ha}(w_i)$ represents the domain-specific correlation score of the word w_i, λ_p and θ_p are the sentiment analysis parameters of the p-th sentiment classifier, λ_q and θ_q are the domain-specific correlation decision parameters of the q-th domain-specific correlation classifier, σ is the activation function, $\vec{\alpha}_p$ and $\vec{\alpha}_q$ are the weight vectors of the sentiment and domain-specific correlation classifiers, respectively. P and Q are the number of classifiers.

After calculating the emotion score and domain-specific relevance score of each word, the analysis results are weighted at multiple levels to obtain the total confidence of each candidate word. The specific formula is as follows:

$$C_{total}(w_i) = \eta_1 \cdot (\frac{F_s(w_i,t) + F_l(w_i+t)}{2}) + \eta_2 \cdot S_{cp}(w_i) + \eta_3 \cdot S_{em}(w_i) - \eta_4 \cdot S_{ha}(w_i) + \eta_5 \cdot \left\|\nabla \vec{v}(w_i)\right\| \tag{11}$$

Where, $C_{total}(w_i)$ represents the comprehensive confidence of the word w_i, η_1, η_2, η_3, η_4, η_5 is the weighting coefficient, which is optimized by training data, and $\left\|\nabla \vec{v}(w_i)\right\|$ represents the gradient norm of the word w_i in vector space. By synthesizing the confidence formula, the results of multi-level and multi-stage analysis can be integrated, and the average of short-term frequency and long-term frequency can be used as the frequency score of words, and then the context compatibility score, emotion score, domain-specific relevance score and gradient norm with the same order of magnitude of words can be combined to obtain the total confidence of each candidate word.

The multi-level confidence threshold is set, and the total confidence is graded. If the comprehensive confidence of the candidate word exceeds the set threshold, it is judged as a new word. The output produces results, including the identified new words and their corresponding confidence scores.

4 Experiment Analysis

4.1 Dataset

In order to realize the recognition of new words in various fields, we collected a large amount of Chinese data from domestic and foreign social platforms and carried out manual annotation. The data set covers 5 fields: Economy and Trade, Social Culture, Science and Technology Education, External Publicity, and Daily Life, with a total of 60,000 pieces of data. The data distribution in specific fields is shown in Table 1.

Table 1. Data in each domain of the dataset

Domain	Quantity
Daily Life	16017
Economy and Trade	10079
Social Culture	14192
Science and Technology Education	10335
External Publicity	9377

The lexicon linked to the disambiguated entities in the dataset comes from the PKUBase Semantic Knowledge Base, which is from the Natural Language Processing and Cognitive Intelligence Laboratory of the Beijing Big Data Institute. The PKUBase Semantic Knowledge Base consists of Chinese RDF semantic knowledge that contains more than 1.3 million entities and more than 13 million knowledge entries.

Table 2 shows partial slice data of the multi-sense disambiguation data set. The first column of the dataset is the label number, the second column is the entity to be disambiguated, the third column is the domain label, and the fourth column is the description of the disambiguated entity, which can be regarded as the entity's righteousness item. It contains entity denotations for 38,629 defined domains. By comparing the entity names in the labeled text with the candidate entities obtained from the tweets, we were able to assess and verify the accuracy of the experimental results. This approach, which is based on real social platform data and has been manually and accurately labeled, ensures that our research can be more relevant to real-world application scenarios, thus improving the efficiency and accuracy of polysemy disambiguation.

Table 2. Partially sliced data from the polysemy disambiguation dataset

Label number	Homonym	Domain	Mean
10	baozhuang (wrap)	Daily Life	Refers to something that wraps merchandise, such as paper, boxes, bottles, etc.
11		Economy and Trade, Social Culture	Refers to the external shaping, beautification and promotion of products, concepts, culture, image, etc. through various means to enhance their attractiveness, competitiveness or influence
32	guang (beam)	Daily Life	A natural phenomenon that refers to the visible light emitted by the sun, light bulbs, etc.
33		Economy and Trade	Means depleted, used up
34		Science and Technology Education	An electromagnetic wave with fluctuating and particle properties

4.2 Evaluation Metric

This paper uses accuracy, recall, and F1 score to measure performance. In single-label classification results, the outcomes can be categorized into True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). Using these four types of results, evaluation metrics for each class can be computed.

This section uses the following common metrics to evaluate the model's performance.

Accuracy (P): Accuracy is the most common, fundamental, and intuitive metric. It is computed using the following formula:

$$P = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

Recall (R): Recall is an important metric that complements accuracy. Recall is used to assess the model's ability to identify TP samples. It is calculated using the following formula:

$$R = \frac{TP}{TP + FN} \tag{13}$$

F1: F1 Score addresses this by combining accuracy and recall into a single metric, providing a comprehensive measure of model performance. The F1 Score is the harmonic mean of precision and recall, with a higher value indicating better model performance. It is calculated using the following formula:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{14}$$

4.3 Experimental Results

4.3.1 High-frequency Word Extraction Experiment

In order to assess the effectiveness of the improved hash-based high-frequency word extraction algorithm proposed in this paper for extracting abnormal high-frequency word quality, a total of 1,000 data from different domains over a continuous period of time are selected as data sources to ensure the diversity and representativeness of the dataset. Also, compared with the traditional TextRank keyword extraction method, the number of keywords extracted from each piece of TextRank data is set to 3 in the experiment.

The experimental results are shown in Table 3. From the indicators in the table, it can be seen that the recall of the improved hash-based high-frequency word extraction method is not as good as that of TextRank, which is due to the fact that the number of keywords extracted by TextRank is more, so the recall rate is high, but the extracted keywords are not in line with the requirements, which leads to a low precision rate. Whereas, the precision rate and F1 value of high-frequency word extraction based on improved hash improved considerably.

Table 3. Performance metrics for extracting anomalous high-frequency words (%)

Method	P	R	F1
TextRank	52.12	68.23	60.35
Our	**68.02**	64.15	**66.13**

4.3.2 Semantic Disambiguation Experiments

In order to evaluate the performance of the proposed experiments, three typical disambiguation methods are selected for comparison, namely Wikify [19], Support Vector Machine (SVM) [20], and Knowledge Base [21]. Wikify is mainly based on predefined entity names and linking rules, and thus lacks real-time dynamic analysis of sentence semantics to determine the specific meaning of polysemy words. The characteristic of disambiguation method based on knowledge base is that it relies heavily on knowledge base, which is usually a relatively static set of knowledge. Therefore, it cannot adapt well when facing dynamic contexts. The SVM disambiguation approach is a graph model combined with entity link disambiguation, and the performance of the model relies on the feature selection of the training data and the model parameters. It is therefore less effective when facing dynamic contextual information. Compared with the above three methods, the proposed method in this paper takes into account the semantic relationships of sentences and paragraphs, and is able to dynamically capture the performance of words in different contexts.

The above methods are tested using a polysemantic

word disambiguation dataset. The results in Table 4 show that the proposed method outperforms the other three disambiguation methods in the polysemy word disambiguation dataset, and the accuracy is improved by 3.59 percentage points compared with Knowledge Base. As can be seen from Figure 3, compared with the other three models, the loss decline curve of this paper's method is smoother, although not as low as SVM at the beginning. It is also much lower than that of the other two models, and the final loss is also very little different from the best result.

Table 4. Comparison of the performance of different methods (%)

Method	P	R	F1
Wikify	71.32	72.59	71.95
SVM	77.46	76.54	77.00
Knowledge Base	81.64	79.21	80.41
Our	**85.23**	**83.63**	**84.42**

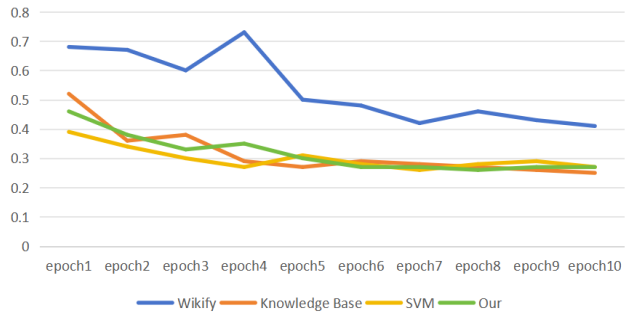

Figure 3. Variation of loss with epoch for different comparison methods (%)

4.3.3 Multi-task Classification Experiment

Table 5. Comparison of the performance of different methods (%)

Dataset	BERT-GCN			ERNIE-GCN		
	P	R	F1	P	R	F1
Daily Life	93.18	90.91	92.03	**94.26**	**93.94**	**94.10**
Economy and Trade	86.63	**86.57**	**86.60**	**86.79**	85.42	86.10
Social Culture	93.28	91.92	92.59	**94.44**	**96.90**	**95.65**
Science and Technology Education	91.08	**93.04**	92.05	**91.98**	92.33	**92.15**
External Publicity	83.19	**86.15**	84.64	**84.53**	85.80	**85.16**
Avg	89.47	89.72	89.58	**90.4**	**90.88**	**90.63**

The performance of different methods is compared in Table 5. BERT-GCN was compared with ERNIE-GCN and analyzed from three perspectives: accuracy, recall, and F1. It concluded that ERNIE-GCN's algorithm has a higher accuracy than BERT-GCN in all domains, especially in Daily life, Social Culture, average recall, and F1. Therefore, this paper adopts ERNIE-GCN as a model for multi-task classification.

4.3.4 New Word Recognition Experiment

This experiment uses a self-constructed dataset to test the neologism recognition algorithm proposed in this paper, and the neologism recognition results on the test set are shown in the following Table 6.

Table 6. Performance metrics for new word recognition algorithms (%)

Method	P	R	F1
Our	85.21	79.51	82.26

4.3.5 Ablation Experiment

In order to evaluate the effectiveness of each module, a series of ablation experiments were conducted and the performance of different configurations was measured by using the F1 score as an evaluation metric, and the results of the experiments are shown in Table 7.

We tested the performance of the algorithms when using only three modules: high-frequency word extraction, semantic disambiguation, and multi-task classification. High-frequency word extraction and multi-task classification have a greater impact on the algorithm performance. In order to explore the synergistic effect between modules, the case of two-by-two feature combination is further investigated, and the results also show that the combination of high-frequency word extraction and multi-task classification has a greater improvement on the algorithm performance. The results of the ablation experiments not only confirm the role of the previously proposed modules, but also emphasize the effectiveness of the multi-model combination strategy.

Table 7. Results of ablation experiments (%)

High-frequency word extraction	Semantic disambiguation	Multi-task classification	F1
√			79.63
	√		78.76
		√	79.12
√	√		80.37
√		√	81.53
√	√	√	82.26

5 Conclusion

In the process of natural language processing, word segmentation is the basic work, and its accuracy has an important impact on the subsequent tasks. However, the effect of word segmentation is often limited by unrecognized new words, which will have a significant impact on subsequent processing steps such as text information extraction and entity recognition. In this paper, we propose a new enhanced new words recognition based on multi-level semantic vectors and multi-task

learning models, which combines the improved hash algorithm, multi-level time series model and contextual semantic graph model to successfully realize the accurate recognition and emotional analysis of new words. First, the unique identifiers of words are calculated by the hash algorithm and a dictionary is generated. The abnormal high-frequency words are identified as potential new word candidates by the comparison of short-term and long-term frequencies. Then, the candidate words are mapped to a high-dimensional vector space to generate a comprehensive semantic vector, and the contextual semantic graph model is used to analyze the context compatibility of the words, which further improves the accuracy of recognition. The multi-task learning emotion analysis model is constructed to calculate emotion scores and domain-specific relevance scores. Finally, by comprehensively calculating each score, the total confidence of each candidate word is obtained, and potential new words related to various fields are identified.

References

[1] H. Abdelhakim, T. Zied, A Hybrid Ensemble Learning Approach for Detecting Bots on Twitter, *International Journal of Performability Engineering*, Vol. 20, No. 10, pp. 610-620, October, 2024. https://doi.org/10.23940/ijpe.24.10.p3.610620

[2] B. B. J. V., J. M. Philip, C. K. T., A. K. P., A Framework for Analyzing the Context of Discussion in Crowd Clusters, *International Journal of Performability Engineering*, Vol. 20, No. 4, pp. 224-231, April, 2024. https://doi.org/10.23940/ijpe.24.04.p4.224231

[3] F. W. Zhai, F. L. He, W. L. Zuo, Chinese word segmentation based on dictionary and statistics, Mini-Micro Systems/ *Journal of Chinese Computer Systems*, Vol. 27, No. 9, pp. 1766-1771, March, 2006. https://doi.org/10.3969/j.issn.1000-1220.2006.09.039

[4] A. Wu, Z. Jiang, Word segmentation in sentence analysis, *Proceedings of the 1998 International Conference on Chinese Information Processing*, Beijing, China, 1998, pp. 169-180.

[5] H. Yu, K. Huang, Y. Wang, D. Huang, Lexicon-augmented cross-domain Chinese word segmentation with graph convolutional network, *Chinese Journal of Electronics*, Vol. 31, No. 5, pp. 949-957, September, 2022. https://doi.org/10.1049/cje.2021.00.363

[6] S. Guo, Y. Huang, B. Huang, L. Yang, C. Zhou, Cwsxlnet: A sentiment analysis model based on Chinese word segmentation information enhancement, *Applied Sciences*, Vol. 13, No. 6, Article No. 4056, March, 2023. https://doi.org/10.3390/app13064056

[7] Y. Shao, Z. Geng, Y. Liu, J. Dai, H. Yan, F. Yang, Z. Li, H. Bao, X. Qiu, Cpt: A pre-trained unbalanced transformer for both Chinese language understanding and generation, *Science China Information Sciences*, Vol. 67, No. 5, Article No. 152102, May, 2024. https://doi.org/10.1007/s11432-021-3536-5

[8] Z. Zhao, Y. Shi, B. Li, Newly-emerging domain word detection method based on syntactic analysis and term vector, *Computer Science*, Vol. 46, No. 6, pp. 29-34, June, 2019. https://doi.org/10.11896/j.issn.1002-137X.2019.06.003

[9] C. Feng, J. An, Neologisms recognition technology based on a variety of statistical indicators, *2022 4th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE*, Shanghai, China, 2022, pp. 170-173. https://doi.org/10.1109/MLBDBI58171.2022.00040

[10] S. Li, A study on the classification of stylistic and formal features in English based on corpus data testing, *PeerJ Computer Science*, Vol. 9, Article No. e1297, April, 2023. http://dx.doi.org/10.7717/peerj-cs.1297

[11] W. Wei, W. Liu, B. Zhang, R. Scherer, R. Damasevicius, Discovery of new words in tax-related fields based on word vector representation, *Journal of Internet Technology*, Vol. 24, No. 4, pp. 923-930, July, 2023. http://dx.doi.org/10.53106/160792642023072404010

[12] J. Duan, M. Wang, Y. Guan, Q. Lin, A method for building Chinese domain lexicon based on new words recommendation, *IEEE 2022 3rd International Conference on Computer Science and Management Technology (ICCSMT)*, Shanghai, China, 2022, pp. 516-522.

[13] Y. Tian, Y. Song, F. Xia, T. Zhang, Y. Wang, Improving Chinese word segmentation with wordhood memory networks, *Proceedings of the 58th annual meeting of the association for computational linguistics*, Online, 2020, pp. 8274-8285. https://doi.org/10.18653/v1/2020.acl-main.734

[14] C. Lin, Y. J. Lin, C. J. Yeh, Y. T. Li, C. Yang, H. Y. Kao, Improving multi-criteria Chinese word segmentation through learning sentence representation, *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023, pp. 12756-12763. https://doi.org/10.18653/v1/2023.findings-emnlp.850

[15] X. Qiu, H. Pei, H. Yan, X. Huang, A concise model for multi-criteria Chinese word segmentation with transformer encoder, *arXiv preprint arXiv: 1906. 12035*, June, 2019. https://arxiv.org/abs/1906.12035

[16] M. Liu, H. Huang, Z. Ding, Pem: A medical named entity recognition method based on proximity enhancement, *IEEE 2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, 2024, pp. 1-8. https://doi.org/10.1109/IJCNN60899.2024.10650135

[17] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv*: 1810. 04805, May, 2019. https://arxiv.org/abs/1810.04805v2

[18] D. Li, R. Zhao, F. Tan, Cwseg: An efficient and general approach to Chinese word segmentation, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023, pp. 1-10. https://doi.org/10.18653/v1/2023.acl-industry.1

[19] B. Hachey, W. Radford, J. Nothman, M. Honnibal, J. R. Curran, Evaluating entity linking with wikipedia, *Artificial intelligence*, Vol. 194, No. 2, pp. 130-150, January, 2013. https://doi.org/10.1016/j.artint.2012.04.005

[20] T. Zhang, K. Liu, J. Zhao, A graph-based similarity measure between Wikipedia concepts and its application in entity linking system, *Journal of Chinese Information Processing*, Vol. 29, No. 2, pp. 58-67, March, 2015.

[21] C. Zhao, H. Y. Li, An entity linking approach for knowledge base question answering, *Journal of Chinese Information Processing*, Vol. 33, No. 11, pp: 125-133, November, 2019.

Biographies

Jin Pan received M.D. degree from Beijing University of Posts and Telecommunications (BUPT) in 2011. He is currently a senior engineer in National Computer Network Emergency Response Technical Team/ Coordination Center of China (CNCERT/CC). His research interests include network security and blockchain.

Yang Chen received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications (BUPT) in 2020. He is currently an engineer in the National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC). His research interests include cryptography, cloud computing and information security.

Chunlu Zhao received M.D. degree from Beijing University of Posts and Telecommunications (BUPT) in 2013. He is currently a senior engineer in National Computer Network Emergency Response Technical Team/ Coordination Center of China (CNCERT/CC). His research interests include cloud computing and artificial intelligence.

Yang Liu received M.D. degree from Beijing University of Posts and Telecommunications (BUPT) in 2009. She is currently a senior engineer in National Computer Network Emergency Response Technical Team/ Coordination Center of China (CNCERT/CC). Her research interests include network security and information security.

Jie Chu graduated from Shandong Technology and Business University. He is currently the assistant director of Institute of Network Technology (Yantai). His main research interests include artificial intelligence and information security.