Postprint: Integrating Semantic Association and BERT for SAO Short Text Classification in Library and Information Science

Authors: Zhang Yujie, Bai Rujiang, Liu Mingyue, Yu Chunliang

Date: 2023-04-01T16:02:54+00:00

Abstract

[Purpose/Significance] To address the issues of semantic feature scarcity and insufficient domain knowledge in SAO-structured short text classification, this study proposes an SAO classification method integrating semantic association and BERT, aiming to enhance short text classification performance. [Method/Process] Utilizing SAO short texts from the library and information science domain as the data source, the study first designs a semantic association scheme comprising three stages—"expansion-reconstruction-denoising"—which extends SAO semantic information through semantic expansion and SAO reconstruction, and addresses noise interference after expansion via semantic denoising; subsequently trains the semantically associated SAO short texts using the BERT model; and finally implements automatic classification in the classification component. [Results/Conclusions] After comparative experiments on different association values, learning rates, and classifiers, the results demonstrate that SAO short text classification achieves optimal performance when the association value is 10 and the learning rate is 4e-5, attaining an average F1 value of 0.852 2. Compared with SVM, LSTM, and BERT alone, the F1 value improves by 0.103 1, 0.153 8, and 0.140 5, respectively.

Full Text

Research on SAO Short Text Classification in Library and Information Science Based on Semantic Association and BERT

Zhang Yujie[1], Bai Rujiang[1], Liu Mingyue[1], Yu Chunliang[2]
[1]Institute of Information Management Research, Shandong University of Technology, Zibo 255049
[2]Yantai University Library, Yantai 264005

Abstract:
[Purpose/Significance] To address the challenges of semantic feature scarcity and insufficient domain knowledge in SAO-structured short text classification, this paper proposes a SAO classification method that integrates semantic association and BERT, aiming to improve short text classification performance. [Method/Process] Using SAO short texts from the library and information science field as the data source, we first designed a semantic association scheme comprising three stages: "expansion-reconstruction-denoising," which extends SAO semantic information through semantic expansion and SAO reconstruction, and resolves noise interference after expansion through semantic denoising. The BERT model was then employed to train the SAO short texts after semantic association, and automatic classification was finally implemented in the classification stage. [Result/Conclusion] After comparing different association values, learning rates, and classifiers, experimental results demonstrate that the SAO short text classification achieves optimal performance when the association value is 10 and the learning rate is 4e-5, with an average F1-score of 0.8522. Compared with SVM, LSTM, and pure BERT, the F1-score improved by 0.1031, 0.1538, and 0.1405, respectively.

Keywords: SAO; short text classification; semantic association; BERT
Classification Number: TP391 G250
DOI: 10.13266/j.issn.0252-3116.2021.16.013

1 Introduction

SAO (Subject-Action-Object) is a specific type of short text structure extracted from papers and patents using particular methods to express key concepts and methodologies [1], representing a fine-grained approach to text expression. SAO consists of three components: Subject (S), Action (A), and Object (O) [2]. Due to its complete grammatical structure, SAO can express richer meanings compared to single keywords and has been widely applied in areas such as potential innovation point mining [3-5], patent feature analysis [6-9], and emerging technology forecasting [10-12].

Research on SAO short text classification facilitates systematic 梳理 of scientific and technological development 脉络 and efficient implementation of text mining tasks. However, current efforts to improve the effectiveness of automatic SAO short text classification still face obstacles, and effectively classifying and organizing the massive yet scattered SAO texts has become an urgent problem to be solved.

Compared with ordinary long documents and short texts, although SAO has a complete grammatical structure, its representational capacity is limited and domain specificity is weak. The only extractable features are Subject and Object and their relationship (Action), making it difficult to obtain effective feature words. Simultaneously, constrained by its expression structure, SAO often suf-

fers from insufficient domain knowledge when applied to domain-specific analysis. Therefore, this paper proposes a SAO short text classification method that integrates semantic association and BERT, using library and information science SAO short texts (hereinafter referred to as LIS SAO) as the data source for empirical validation. The aim is to enrich the representational capacity of SAO short texts, thereby addressing the problems of semantic feature scarcity and insufficient domain knowledge in SAO classification and improving classification performance.

2 Related Research

The essence of SAO is a triple (Subject-Action-Object) consisting of two nodes and their relationship, forming the basic element for knowledge graph construction. Similar structures include SPO (Subject-Predicate-Object) and SVO (Subject-Verb-Object) [6]. These concepts all identify syntactic structures and dependency relationships in sentences through entity recognition and syntactic analysis to extract subject-verb-object elements. The differences between SAO, SPO, and SVO lie in their application scenarios and domains, with varying emphases on predicate selection. In recent years, SAO has found extensive applications in knowledge mining and discovery, potential innovation point mining, and patent feature analysis.

Hu Zhengyin et al. constructed a semantic TRIZ methodology, process, and key technologies at the micro-level of SAO, using large-aperture optical component patents as an example to build domain-specific semantic TRIZ, demonstrating that the proposed method could effectively achieve semi-automatic construction of domain-specific semantic TRIZ [13]. Subsequently, the same author integrated SAO triples with simple knowledge objects and bidirectional Transformer [25] encoders, designing numerous multi-head attention mechanisms to construct fine-grained, multi-dimensional domain technology indexes, achieving domain knowledge prism, TRIZ-oriented semantic retrieval, and patent visualization analysis functions [7]. Wang Xuefeng et al. [3] proposed a method for identifying R&D partners based on solution similarity through SAO triple analysis, using the solar cell industry as a case study. The results showed that this approach could help companies understand relationships between research targets. Later in 2019, they proposed research on selecting innovative solutions based on SAO structure [5], which started from specific research problems in the target field, searched for potential solutions across all domains, and evaluated these solutions from both technical feasibility and expected effectiveness perspectives, demonstrating the method's validity.

SAO classification can be viewed as a special type of text classification in Natural Language Processing (NLP)—namely, short text classification—which essentially converts text content into machine-recognizable vectorized representations and automatically learns text features to identify different categories through

machine learning. However, unlike long document classification, short texts are characterized by limited word count and sparse features, making it difficult for conventional methods to capture effective features. Consequently, most scholars have focused research on machine learning, deep learning, hybrid models, pre-trained models, and data augmentation.

Traditional machine learning methods for short text classification, such as Support Vector Machines [14], Bayesian models [15], Hidden Markov Models [16], and Random Forests [17], construct feature engineering from sample data before feeding it into specific classifiers for training and prediction [18-19]. However, these methods require extensive feature engineering in the early stage, have weak generalization capabilities, and cannot fully utilize large-scale data for feature learning [20]. Deep learning methods for short text classification overcome these limitations by focusing on model construction and parameter tuning, fitting feature values through deep non-linear transformations on large training datasets. Deep learning methods such as Recurrent Neural Networks [21] and their variants [22-24], Convolutional Neural Networks [21], and attention mechanisms [25] have all demonstrated good performance in short text classification. Deng Sanhong et al. [26] fused Long Short-Term Memory networks with character embedding methods for Chinese book tag classification, training models on short text features such as titles and subject terms, achieving good results in classification experiments on five categories of bibliographic data from three universities. Zhao Yajuan [27] and Francck [28] respectively utilized RNNs, CNNs, and hybrid methods for short text classification in patents and dialogue act domains. Zhang Chengzhi [29], Tao Zhiyong [30], and Yu Bengong [31] improved or fused hierarchical attention networks, providing many research ideas for short text feature representation. However, these methods lack divergence of deep text meanings and require large amounts of labeled data for training, imposing high demands on both data quantity and quality; simple, small-scale data cannot adapt to complex network models.

In 2018, Google proposed the BERT [32] pre-trained model, which employs multiple bidirectional Transformer encoders, designs numerous multi-head attention mechanisms, and learns general knowledge from large-scale training data, supplemented by fine-tuning with small amounts of domain data, achieving state-of-the-art (SOTA) results on multiple downstream tasks including text classification. X. Qiu et al. [33] comprehensively compared various BERT-based text classification methods, proposing many ideas for fine-tuning strategies, further pre-training, and multi-task training. J. S. Lee [34] and X. Lu [35] used BERT for fine-tuning on patent data classification, both achieving satisfactory results.

In research on integrating semantic information and domain knowledge into BERT, some scholars have improved BERT's input mode to enhance text semantic information [36]. W. Liu et al. [37] proposed K-BERT, which maps training data to domain knowledge triples to increase domain knowledge in input data while adding a visual layer to solve knowledge noise problems, achieving good performance on multiple datasets. S. Yu et al. [38] proposed constructing

auxiliary sentences and domain knowledge for texts, converting classification tasks into binary sentence pairs, and investigated the impact of learning strategies, learning rates, sequence length, and hidden state vectors on classification results.

The aforementioned studies provide important insights for this paper: relying on pre-trained models and conducting domain-specific semantic association on datasets can alleviate the problems of semantic feature scarcity and insufficient domain knowledge in short texts. However, for the special SAO-structured short texts, particularly in specialized domains such as library and information science, no relevant studies have been conducted to validate or propose ideal solutions. Building upon these studies, this paper further investigates SAO short text classification.

---

3 Methodology for SAO Short Text Classification Integrating Semantic Association and BERT

The research framework of this paper is shown in Figure 1 [Figure 1: see original paper], comprising three main components: semantic association, BERT integrated with semantic association, and classification. Semantic association aims to enhance SAO semantic representation capability and resolve semantic noise interference, BERT is used to fine-tune SAO data after semantic association, and appropriate classifiers are selected in the classification stage to achieve automatic short text classification. Additionally, the paper compares the effects of different models, association values, learning rates, and classifiers on short text classification results to explore the optimal solution for LIS SAO short text classification.

3.1 Semantic Association

The semantic association scheme proposed in this paper consists of three parts: semantic expansion, SAO reconstruction, and semantic denoising. Its purpose is to extend more contextual information for SAO, capture more domain knowledge during feature encoding, and simultaneously prevent semantic expression from deviating from the original SAO due to excessive association. Therefore, the scheme includes three stages: "expansion-reconstruction-denoising." Let the input be SAO, and the trained Word2Vec [40] model in the library and information science field be denoted as M.

3.1.1 Semantic Expansion Compared with general SAO, LIS SAO faces not only semantic feature scarcity but also insufficient domain knowledge. Specifically, short text automatic classification often encounters challenges such as scarce publicly labeled samples in the library and information science field, high manual annotation costs, inconsistent annotation quality due to varying knowledge levels, non-significant classification features within the domain, and

sparsity of domain encoding. Therefore, the purpose of semantic expansion is to extend more LIS domain contextual information for SAO and capture more domain knowledge during feature encoding. To this end, we trained a Word2Vec model for the library and information science field, which adapts the most similar synonymous expressions in vector space for Subject and Object respectively, enriching SAO's expressive capability while supplementing additional domain knowledge.

During semantic expansion, the most similar candidate synonymous expressions are returned by calculating cosine distance [39], which reflects the similarity of two words' spatial positions. The calculation formula is shown in Equation (1), where X is the target word and Y is all words in the model space. The candidate words calculated are mapped with the original SAO and mounted under their corresponding query terms to generate the expanded SAO short text, stored as an SAO tree in a tree structure, denoted as T. Its calculation formula is shown in Equation (2), where n is the association value.

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}}$$

$$T = \{(S \dots S_n)A(O \dots O_n)\}$$

The core of LIS SAO short text lies in Subject and Object. Action, as the predicate, only reflects the subject-object relationship of SAO; therefore, we do not expand Action but only expand Subject and Object. For example, after performing n=2 expansion on "university library, construct, learning commons," the structure is shown in Figure 2(a).

3.1.2 SAO Reconstruction The BERT model accepts sequential structure input, so the SAO tree needs to be constructed into linear sequence-structured text, denoted as L. There are two reconstruction schemes at this point, using "university library, construct, learning commons" as an example.

Scheme 1 is {university library, university library, research university library, construct, learning commons, information commons, physical space}, shown in Figure 2(b), denoted as L1, expressed by Equation (3):

$$L1 = \{S_0 \dots S_n AO_0 \dots O_n\}$$

Scheme 2 is {university library construct learning commons, university library construct information commons, university library construct physical space, university library construct information commons, university library construct learning commons, university library construct physical space, research university library construct physical space, research university library construct

learning commons, research university library construct information commons}, shown in Figure 2(c), denoted as L2, expressed by Equation (4):

$$L2 = [S_0 \dots S_n] \times A \times [O_0 \dots O_n] = \{S_0AO_0, \dots, S_nAO_n\}$$

Both reconstructed SAOs have the problem of deviating from the original SAO's meaning, i.e., semantic noise. The semantic noise of L1 lies in losing grammatical structure relationships; the original subject-predicate-object relationship cannot express complete semantics after reconstruction, causing misalignment in encoded word relationships and leading to incorrect contextual information. The semantic noise of L2 lies in overloaded SAO combinations after expansion, causing excessive association and resulting in expanded outcomes that contradict the original meaning. To address this problem, this paper proposes a semantic denoising solution.

3.1.3 Semantic Denoising Due to the differential distribution of training corpora and the black-box nature of word vector representation, Word2Vec inevitably introduces irrelevant or even contradictory feature words for LIS SAO semantic expansion. Therefore, expanded and reconstructed LIS SAOs require further "cleansing." The purpose of semantic denoising is to reduce the noise interference of SAOs after semantic association on the original SAO while maximizing the retention of associative information. Based on this, we borrow the core idea of the attention mechanism to "score" each expanded word, selectively mounting or forgetting SAOs after semantic association to highlight key points and discard redundancy.

The specific approach is: assign weights to each expanded word, with weight values represented by Word2Vec similarity between the target word and retrieved words. Different weights represent the importance of different expanded words, with original SAO components each having a weight of 1. Then, perform weighted summation on each SAO in L2, sort them, and take the top n+1 SAOs as the SAOs after semantic association, i.e., L, expressed by Equation (5), where w represents the different weights assigned to different expanded words.

$$L = \{\sum wS_0wAwO_0, \dots, \sum wS_nwAwO_n\}$$

As shown in Figure 2(d), after semantic denoising, "university library, construct, learning commons" is expressed as {university library construct learning commons (3), university library construct information commons (2.781), university library construct learning commons (2.735), university library construct physical space (2.717)}. The denoised SAO reduces noise interference based on semantic association, maximally ensures semantic integrity and divergence, retains the positional relationships of SAO structure, and guarantees the interpretability of position embedding during BERT word embedding.

3.2 BERT Integrated with Semantic Association

After semantic association, SAO is input into BERT for fine-tuning and training. This paper selects BERT because, as a pre-trained model, it already carries general domain prior information based on large-scale data training. Using BERT for LIS SAO classification only requires fine-tuning SAO data after semantic association, thereby alleviating overfitting or underfitting problems caused by retraining complex model parameters from scratch, with the aim of improving the representation capability of LIS SAO word vectors.

BERT pre-trained parameters combined with SAO after semantic association obtain new training parameters, which sequentially pass through word embedding and multi-layer bidirectional Transformers. Word embedding converts input text into vector representations. Transformers capture text weight information through encoders.

Word embedding mainly consists of three processes: Token Embedding, Segment Embedding, and Position Embedding, as shown in Figure 3. Token Embedding converts SAO into character-level one-dimensional vector representations through BERT's character lookup table, randomly masking some characters during MASK to obtain bidirectional information from left to right and right to left, with [CLS] marking the beginning of an SAO and [SEP] marking the end. Segment Embedding marks different symbols to obtain global semantic information of the text and identify different SAOs, fusing with character-level vectors. Position Embedding marks contextual information before and after. Finally, token embedding, segment embedding, and position embedding are summed to output the final word embedding representation.

BERT's position embedding is an important feature that distinguishes it from other models. Position information makes each character's impact on other characters not completely identical, enabling BERT to dynamically capture associations before and after words based on context. During position embedding, character weight coefficients are not determined by fixed parameters but are fused with the current character's weight after the previous character calculates its weight. When generating the next character, the originally fixed parameter w is replaced by a dynamically changing parameter w_i based on the previous word, injecting previous character information into each character of the SAO. The longer the input, the more important the weight coefficient.

As shown in Figure 3(a), the original SAO can obtain insufficient contextual information, with expansion weight coefficients only calculable up to 15 when computing contextual information. The word embedding of SAO after semantic association is shown in Figure 3(b). After association, the weight coefficient is higher than that of the original SAO, placing greater emphasis on embedding of contextual relationships, allowing the model to capture richer detailed information.

After word embedding, it connects to Transformer encoders. As BERT's feature

extractor, Transformer uses a multi-layer bidirectional structure to calculate hidden state vectors, containing 12 Transformer encoders and multi-head attention, which accumulate layer by layer to form BERT. Transformer encoders incorporate position information into encoding through multi-head attention, considering the weight influence of the previous word on the current word, aligning input dimensions with output dimensions, and outputting hidden states after normalization.

3.3 Classification

After BERT training, SAO is represented in vector form. In the classification stage, corresponding classifiers are connected to achieve automatic classification. For short text multi-classification, a fully connected layer or other pre-built network models are typically connected, with Softmax used as the activation function for classification.

The process of Softmax achieving final classification prediction results is as follows: Given a set D containing i SAOs:

$$D = \{(L_1, label_1), \dots, (L_i, label_i)\}$$

where L_i represents the i-th SAO after semantic association, i R, and label_i represents the category corresponding to the i-th SAO, label_i {1, 2, …, c}, with c being the number of categories. For any SAO, the goal is to calculate the conditional probability distribution through the Softmax function, i.e., the probability that the SAO belongs to each category, returning a c-dimensional matrix where the category with the maximum probability value is the category to which the SAO belongs.

After the BERT layer training, model parameters and hidden state vector H are output. Therefore, the Softmax objective is to calculate the probability distribution P(label_i|H[CLS],), as shown in Equation (7). After the above strategy, the classification probability value for input SAO is obtained.

$$P(label_i|H[CLS], L) = \frac{e^{label_i|H[CLS],L}}{\sum_{j=1}^{c} e^{label_j|H[CLS],L}}$$

3.4 Evaluation Metrics

To evaluate SAO short text classification effectiveness, this paper adopts Precision, Recall, and F1-score as evaluation metrics, as shown in Equations (8) to (10). The P-value is commonly used to assess the proportion of correct predictions; a higher P-value indicates better prediction accuracy and model performance. The R-value measures classification accuracy; a higher R-value indicates better model performance. Typically, precision and recall cannot simultaneously achieve high standards, and using either P-value or R-value alone

as a measurement metric lacks comprehensiveness. Therefore, the F1-score uses weighted harmonic mean.

$$P = \frac{\text{Predicted Correct Results}}{\text{All Predicted Results}}$$

$$R = \frac{\text{Predicted Correct Results}}{\text{All Results in Sample}}$$

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

---

4 Empirical Research

Based on the above design 思路, this section conducts empirical research. To compare the differences between the proposed method integrating semantic association and BERT with traditional machine learning and deep learning approaches, the experiments select Support Vector Machine (SVM) and Long Short-Term Memory network (LSTM) as baseline models. To compare the impact of different numbers of expanded words on classification effectiveness, experiments with different association values are conducted. To compare the influence of different learning rates and classifiers on results, comparative experiments on different learning rates and classifiers are performed based on the group with optimal association value classification performance.

4.1 Experimental Environment

Hardware configuration: Intel E5-2609v4 + NVIDIA TESLA P4$×$1
Software configuration: Win10 + Python 3.6 + TensorFlow 1.5 + Keras 2.1 + PaddlePaddle 1.7

4.2 Corpus Sources and Datasets

4.2.1 Corpus Sources The data in this paper originates from 18 CSSCI journals in the library and information science field, including *Journal of Library Science in China*, *Journal of Intelligence*, *Journal of Academic Libraries*, *Library and Information Service Knowledge*, *Library and Information*, *Information and Documentation Services*, *Library and Information Work*, *Information Studies: Theory & Application*, *Journal of Intelligence*, *Information Science*, *Library Tribune*, *Journal of the National Library of China*, *Data Analysis and Knowledge Discovery*, *New Technology of Library and Information Service*, *Library Science Research*, *Library*, *Library Development*, *Library Journal*, and *Modern Intelligence*. The top 500 most-cited papers from each journal were selected, comprising 9,000 bibliographic records from 2000-2018, with each record containing attributes such as title and author keywords.

4.2.2 Word2Vec Dataset and Model The Word2Vec dataset is used to train the Word2Vec model for subsequent semantic association of LIS SAO. To improve the quality and novelty of associative words, the training dataset added 11,931 bibliographic records from four journals (*Journal of Library Science in China*, *Library and Information Work*, *Journal of Intelligence*, and *Data Analysis and Knowledge Discovery*) published between 2000-2020, based on the aforementioned 9,000 records. Since the data attributes differed from the original 9,000 records and the model construction does not require vocabulary deduplication, the final dataset for building the Word2Vec model totaled 20,931 entries.

Before Word2Vec training, preprocessing operations such as word segmentation, stop word removal, case conversion, and deletion of useless symbols are required. To maximize model quality, we extracted 60,503 entries from the *Encyclopedia of China: Library Science, Information Science, and Archival Science* [41] and *New Dictionary of Library and Information Science* [42] as an external dictionary for jieba word segmentation, defined 4,652 common characters, words, and symbols as a stop word list, and used the Gensim library for training. Parameters were set as: dimension 100, dictionary pruning count 3, training algorithm Skip-gram, skip window 5, with training completed after 10 iterations. Visualized word vector results for querying words similar to "information" are shown in Figure 4 [Figure 4: see original paper].

4.2.3 SAO Short Text Classification Dataset The library and information science SAO short text classification dataset is the target data for automatic classification in this paper. The extracted fields from paper data include title and abstract. The extraction method is based on Harbin Institute of Technology's LTP dependency parsing and semantic role labeling open-source project for SAO triple extraction. The difference is that this study uses a custom LIS dictionary and stop word list during word segmentation. After calling the TripleExtractor class method in the program, the program functionality was rewritten and applied to our data. Additionally, after extraction, we defined SAO filtering and cleaning rules [43] to filter SAO quality and ensure usability.

After cleaning, the manually annotated library and information science SAO short text data totaled 11,021 entries, each containing four attributes: subject, action, object, and label. The character length distribution and word frequency distribution of SAO short texts are shown in Figure 5 [Figure 5: see original paper].

Classification labels reference the eight categories published by the National Technical Committee for Standardization in *Library, Information and Documentation Terms 2019* [44]. After consulting relevant literature, standards, patents, and multiple expert discussions, they were finalized as six major categories: information resource construction, information organization, library and information work management, information service and user research, intelligence analysis and research, and others. After manual annotation, domain experts provided feedback, and the final confirmation was made after multiple

discussions and revisions by four LIS experts and scholars. The correspondence between category names, IDs, and quantities is shown in Table 1 . The above data were randomly split in an 8:2 ratio, with random seeds set to ensure controllability of random sampling.

4.3 SAO Short Text Classification Experiments Based on Baseline Models

4.3.1 SAO Short Text Classification Experiment Based on SVM Support Vector Machine (SVM) is one of the most widely used machine learning algorithms. Its principle is to find a hyperplane in N-dimensional space to partition data points, maximizing the distance between the two categories and the plane. Relevant studies [45-47] using SVM models have achieved good results, so this paper selects SVM as one of the baseline models. SVM constructs data features after preprocessing such as word segmentation and stop word removal, sequentially undergoes feature selection and feature weight calculation, constructs an SVM classifier, and after multiple iterations and optimization, achieves an accuracy of 0.75 and an average F1-score of 0.7491. Parameters for each category are shown in Table 2 .

4.3.2 SAO Short Text Classification Experiment Based on LSTM Long Short-Term Memory network (LSTM) is a variant of Recurrent Neural Network (RNN). With its gate mechanism, it reduces long-term dependencies in sentences, effectively mitigating gradient vanishing and explosion problems, and is widely used in text classification tasks [26, 48], thus selected as a baseline model in this paper. LSTM input encoding maps to a dictionary that assigns an ID to each word before vectorization, converting each SAO into a vector of integer sequences. The activation function is set to Softmax, and the loss function is set to categorical cross-entropy. After multiple training iterations, the optimal effect is achieved when Epochs=10 and Batch_{size}=32, with an accuracy of 0.70 and an average F1-score of 0.6984. 各项指标如表 3 所示:

4.4 SAO Short Text Classification Experiments Integrating Semantic Association and BERT

4.4.1 SAO Short Text Classification Experiments with Different Association Values Different numbers of expanded words can generate SAOs of different lengths and semantics. To compare the impact of association values on classification effectiveness, this section conducts experiments with different association values. Considering hardware conditions and Word2Vec model size, association value n was set to 0, 5, 10, and 15 respectively. The Chinese version of BERT was selected as the model. Four experiments adopted unified configuration parameters: Epochs=10, Batch_{size}=32, transfer optimization strategy using PaddlePaddle's encapsulated AdamWeightDecayStrategy, Weight_{decay}=0.01, Warmup proportion=0.1, optimizer=Adam, learning rate uniformly set to 4e-5, classifier set to fully connected network, using

Softmax activation function. Experimental data has the text content in the first column and text category in the second column, separated by Tab keys in tsv format input. After training, 各项指标如表 4 所示:

Observation shows that when n=10, the average F1-score reaches 0.8073, achieving optimal performance, indicating that SAO semantic expression capability is best when 10 words are expanded. Subsequent comparative experiments will be based on n=10.

4.4.2 SAO Short Text Classification Experiments with Different Learning Rates Learning rate is an important factor affecting classification metrics. Different learning rates have different impacts on loss values during training: too large a learning rate can easily cause gradient explosion, making it difficult to smooth loss step amplitudes and preventing model convergence; too small a learning rate leads to slow convergence speed, causing data overfitting. This section conducts comparative experiments with learning rates (using scientific notation) set to 1e-6, 2e-5, 4e-4, and 4e-5 respectively, based on n=10, with other configuration parameters unchanged. After training, 各项指标如表 5 所示:

4.4.3 SAO Short Text Classification Experiments with Different Classifiers BERT classification tasks typically use a simple fully connected network as the classifier, with Softmax as the activation function for automatic classification. For SAO texts after association where multiple sentences express similar meanings, can selecting other network models as classifiers improve classification effectiveness? This section compares the impact of using fully connected network (FullyConnectedNetwork, FC) and LSTM network as classifiers on P-value, R-value, and F1-score, based on n=10 and learning_{rate}=4e-5. The fully connected network accepts sentence-level features, outputting the corresponding vector for [CLS] in the format [-1, emb_{size}]; when set to LSTM, it outputs character-level features with structure [-1, max_{{{seq}}*{{len}}}, *emb*{size}]. When changing the classifier, only one layer of network needs to be added in Task. After training, 各项指标见表 6 :

4.5 Results Analysis

4.5.1 Comparative Analysis of Different Classification Model Results The average P-value, R-value, and F1-score across different models are shown in Figure 6 [Figure 6: see original paper], where BERT takes the group with the highest results for comparison. Through comparison, it can be found that the average F1-score of SAO after integrating semantic association and BERT is higher than SVM and LSTM, at 0.8524, 0.8531, and 0.8522 respectively. As shown in Figures 7(a) and 7(b), BERT achieves the highest F1-score across all categories, with category distribution in boxplots remaining relatively stable. This is because BERT, based on large-scale pre-trained corpora and combined with the semantic association scheme proposed in this paper, uses a fusion of

general and domain knowledge to more explicitly represent semantic information. Simultaneously, semantic denoising can forget associative words with low relevance and high noise, further improving completeness and thus achieving better classification results.

Additionally, SVM shows better recognition performance than LSTM, with an F1-score reaching 0.7491. As a traditional machine learning algorithm, SVM can effectively handle high-dimensional feature samples and perform feature extraction with small sample sizes. Compared with complex deep learning models, it does not completely rely on parameter features and has stronger generalization ability in single domains. In contrast, the deep learning model LSTM performs worse than other models for data classification. As a variant of RNN, LSTM has more complex model parameters and computational requirements, needing large amounts of data to learn feature differences between different categories. For experiments with insufficient data, it easily loses encoding information. SAO short text structure is simple, does not require learning overly long sequences, has relatively fixed inter-word quantities, and the constructed dictionary scale is not large, thus cannot leverage the advantages of large-scale parameter computation, leading to inferior classification results.

4.5.2 Comparative Analysis of Different Association Value Results Classification effectiveness varies under different association values. As shown in Figure 7(c), the F1-score is significantly affected by n. When n=0 (no semantic association performed), classification performance is worst, maintained between 0.65-0.75. When n=5, the overall F1-score increases, indicating that expanding 5 words yields better results than no expansion. When n=10, F1-scores across categories improve significantly, maintained around 0.8. As shown in Figure 7(d), when n=10, differences between categories are least significant (i.e., smallest gaps) and stability is best. However, when n reaches 15, F1-scores drop sharply to between 0.67-0.75, with large differences between maximum and average F1-scores, poor classification effectiveness, and relatively poor stability. This is because the Word2Vec training data scale is limited and cannot fully match the most similar expressions for each word. Therefore, when n=15, the overall relevance of associative words decreases, causing increased deviation between SAO texts and their corresponding categories and reduced classification effectiveness.

From 0 to 5 to 10, as the association value increases, the classification F1-score improves accordingly, indicating that with more expanded words, SAO short text semantic information becomes richer and differences between categories become more significant. When the association value reaches 15, performance declines, suggesting that SAO short text classification effectiveness improves with association value but needs to be controlled within a local range.

4.5.3 Comparative Analysis of Different Learning Rate and Classifier Results Changes in loss and accuracy under different learning rates are shown

in Figure 8 [Figure 8: see original paper]. When learning rate is set to 1e-6 and 4e-4, the loss during training decreases slowly, with maximum accuracy around 0.6. As learning rate decreases, training loss gradually declines and accuracy gradually improves. When learning rate is 4e-5, optimal performance across four metrics is achieved. Category F1-scores and their distribution under this setting are shown in Figures 7(e) and 7(f). F1-scores are most concentrated and stable at 2e-5 and 4e-5, while others are more dispersed with higher F1-score dispersion. It can be seen that classification effectiveness across LIS SAO categories has a relatively close relationship with learning rate size; smaller learning rates yield higher classification metrics and better dataset performance, which aligns with the general characteristics of BERT and other deep learning models.

Differences between various category metrics under different classifiers, using fully connected layer and LSTM layer after BERT, are significant in Figures 7(g) and 7(h). Using only the fully connected layer achieves an average F1-score of 0.8522, with category F1-scores maintained around 0.85 overall. However, connecting LSTM before using Softmax activation function for classification yields only 0.7602, similar to the baseline LSTM. The vector calculation of connecting LSTM after BERT for classification does not leverage the advantages of increased parameters; the effect is 反而不如直接连接的全连接网络。

In summary, after comparing the effects of different models, association values, learning rates, and classifiers on SAO classification, the results demonstrate that compared with traditional machine learning and deep learning, the SAO short text classification method integrating semantic association and BERT has more significant advantages. Compared with pure SVM, LSTM, and BERT classification models, the F1-score improved by 0.1031, 0.1538, and 0.1405 respectively. Classification effectiveness shows a positive correlation with the number of associative words within a local range. With a fixed association value, learning rate and classifier also have certain impacts on results. Ultimately, when the association value is 10 and the learning rate is 4e-5, SAO classification performance reaches optimization.

However, the method proposed in this paper still has certain limitations. Due to the limited corpus of the semantic association model, experiments cannot perform sufficient semantic association for each SAO short text, resulting in decreased classification effectiveness when the association value is further increased. Additionally, this method has not been extended to more domains for adaptability testing. In future research, we will further investigate these issues.

References

[1] CASCINI G, FANTECHI A, SPINICCI E. Natural language processing of patents and technical documentation[C]//International workshop on documentation analysis systems. Berlin: Springer, 2004: 508-520.

[2] CHOI S, PARK H, KANG D, et al. An sao-based text mining approach to building a technology tree for technology planning[J]. Expert systems with applications, 2012, 39(13): 11443-11455.
[3] WANG X, WANG Z, HUANG Y, et al. Identifying r&d partners through subject-action-object semantic analysis in a problem & solution pattern[J]. Technology analysis & strategic management, 2017, 29(10): 1167-1180.
[4] TSOU RIKOV V M, BATCHILOLS, SOVPELIV. Document semantic analysis/selection with knowledge creativity capability utilizing subject-action-object (sao) structures: U.S. Patent 6,167,370[P]. 2000-12-26.
[5] Fu Yun, Wang Xuefeng, Li Jia, et al. Research on innovative solution selection based on SAO structure—taking air purification technology as an example[J]. Library and Information Service, 2019, 63(6): 75-84.
[6] Xu Haiyun, Wang Zhenmeng, Hu Zhengyin, et al. A review of key technologies for identifying technology themes using patent text analysis[J]. Information Studies: Theory & Application, 2016, 39(11): 131-137.
[7] Hu Zhengyin, Liu Chunjiang, Wei Ling, et al. Design and practice of domain patent technology mining system oriented to TRIZ[J]. Library and Information Service, 2017, 61(1): 117-126.
[8] Yang Chao, Zhu Donghua, Wang Xuefeng, et al. Patent technology theme analysis: LDA topic model method based on SAO structure[J]. Library and Information Service, 2017, 61(3): 86-94.
[9] CHANG P L, WU C C, LEU H J. Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display[J]. Scientometrics, 2010, 82(1): 5-19.
[10] GUO J, WANG X, LI Q, et al. Subject-action-object-based morphology analysis for determining the direction of technological change[J]. Technological forecasting and social change, 2016, 105: 27-40.
[11] LI X, WANG J J, YANG Z. Identifying emerging technologies based on subject-action-object[J]. Journal of intelligence, 2016, 35(3): 80-84.
[12] Wang Xiaoyu, Miao Hong, Wang Fang. Identification of cross-domain application of technological knowledge and potential technical solutions[J]. Library and Information Service, 2016, 60(23): 87-96.
[13] Hu Zhengyin, Fang Shu, Zhang Xian, et al. Research on personalized semantic TRIZ construction[J]. Library and Information Service, 2015, 59(7): 123-131.
[14] MANEKA S, SHENOY P D, MOHAN M C, et al. Aspect term extraction for sentiment analysis in large movie reviews using gini index feature selection method and SVM classifier[J]. World Wide Web, 2017, 20(2): 135-154.
[15] BACHHETY S, DHINGRA S, JAIN R, et al. Improved multinomial bayes approach for sentiment analysis on social media using gini index feature selection method and SVM classifier[J]. World Wide Web, 2017, 20(2): 135-154.
[16] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.
[17] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
[18] Gao Jinyong, Xu Chaojun, Feng Yixuan. Application of iterative TFIDF

in short text classification[J]. Information Studies: Theory & Application, 2011, 34(6): 120-122.
[19] Fan Yunjie, Liu Huailiang. Research on Chinese short text classification based on Wikipedia[J]. New Technology of Library and Information Service, 2012(3): 47-52.
[20] MINAEE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning-based text classification: a comprehensive review[J]. arXiv preprint arXiv:2004.03705, 2020.
[21] YIN W, KANN K, YU M, et al. Comparative study of cnn and rnn for natural language processing[J]. arXiv preprint arXiv:1702.01923, 2017.
[22] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
[23] OLAH C. Understanding lstm networks[EB/OL][2015-8-27]. https://colah.github.io/posts/2015-08-Understanding-LSTM/
[24] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.
[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[26] Deng Sanhong, Fu Yangyangzi, Wang Hao. Chinese book multi-label classification based on LSTM model[J]. Library and Information Service, 2012, 56(9): 114-119.
[27] Lü Lucheng, Han Tao, Zhou Jian, et al. Research on automatic Chinese patent classification method based on deep learning[J]. Library and Information Service, 2020, 64(10): 75-85.
[28] LEE J, DERNONCOURT F. Sequential short-text classification with recurrent and convolutional neural networks[J]. arXiv preprint arXiv:1603.03827, 2016.
[29] Qin Chenglei, Zhang Chengzhi. Academic text structure function recognition based on hierarchical attention network model[J]. Data Analysis and Knowledge Discovery, 2020, 4(11): 26-42.
[30] Tao Zhiyong, Li Xiaobing, Liu Ying, et al. Improved attention short text classification method based on bidirectional LSTM network[J]. Data Analysis and Knowledge Discovery, 2019, 3(12): 21-29.
[31] Yu Bengong, Zhu Mengdi. Research on question classification based on hierarchical attention multi-channel convolutional bidirectional GRU[J]. Data Analysis and Knowledge Discovery, 2020, 4(8): 50-62.
[32] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[33] SUN C, QIU X, XU Y, et al. How to fine-tune bert for text classification?[C]//China national conference on Chinese computational linguistics. Cham: Springer, 2019: 194-206.
[34] LEE J S, HSIANG J. Patentbert: Patent classification with fine-tuning a pre-trained bert model[J]. arXiv preprint arXiv:1906.02124, 2019.
[35] LU X, NI B. BERT-CNN: A hierarchical patent classifier based on a

pre-trained language model[J]. arXiv preprint arXiv:1911.06241, 2019.
[36] Liu Huan, Zhang Zhixiong, Wang Yufei. Research review on main optimization and improvement methods of BERT model[J/OL]. Data Analysis and Knowledge Discovery: 1-17[2021-01-05]. https://doi.org/10.11925/infotech.2096-3467.2020.0965.
[37] LIU W, ZHOU P, ZHAO Z, et al. K-BERT: Enabling language representation with knowledge graph[J]. arXiv preprint arXiv:1909.07606, 2019.
[38] YU S, SU J, LUO D. Improving BERT-based text classification with auxiliary sentence and domain knowledge[J]. IEEE access, 2019, 7: 176600-176612.
[39] ORNKHO L K, YANG W. Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet[J]. Future Internet, 2019, 11(5): 114.
[40] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[41] Editorial Committee of Encyclopedia of China. Encyclopedia of China: Library Science, Information Science, and Archival Science[M]. Beijing: Encyclopedia of China Publishing House, 2002.
[42] Qiu Dongjiang. New Dictionary of Library and Information Science[M]. Beijing: Scientific and Technical Documentation Press, 2006.
[43] Bai Rujiang, Zhang Qingzhi, Sun Yigang. Research on knowledge gene expression and inheritance and variation of scientific and technological literature[J]. Library and Information Service, 2020, 64(4): 78-87.
[44] Library, Information and Documentation Terminology Committee. Library, Information and Documentation Terms 2019[M]. Beijing: Science Press, 2019.
[45] ASHKAN J, HAMED E, MIHAN H, et al. Improvement in automatic classification of Persian documents by means of support vector machine and representative vector[C]//International conference on innovative computing technology. Berlin: Springer, 2011: 282-287.
[46] Yang Min, Gu Jun. Research on automatic Chinese bibliography classification based on SVM and its application[J]. Library and Information Service, 2012, 56(9): 114-119.
[47] Wang Dongbo, He Lin, Huang Shuiqing. Research on automatic classification of pre-Qin philosophical classics based on support vector machine[J]. Library and Information Service, 2017, 61(12): 71-76.
[48] WANG J H, LIU T W, LUO X, et al. An LSTM approach to short text sentiment classification with word embeddings[C]//Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018). Hsinchu: ACLCLP, 2018: 214-223.

Author Contributions

Zhang Yujie: Experiment implementation, paper writing;
Bai Rujiang: Research topic selection and design, paper revision and review;

Liu Mingyue: Experimental scheme design, paper revision;
Yu Chunliang: Data corpus processing, paper revision.

English Abstract

Research on SAO Short Text Classification in LIS Based on Semantic Association and BERT

Zhang Yujie[1], Bai Rujiang[1], Liu Mingyue[1], Yu Chunliang[2]
[1]Institute of Information Management Research, Shandong University of Technology, Zibo 255049
[2]Yantai University Library, Yantai 264005

Abstract: [Purpose/significance] Aiming at the shortage of semantic features and insufficient domain knowledge in the classification of SAO structure short texts, this paper proposes a SAO classification method combining semantic association and BERT in order to improve the classification effect of short texts. [Method/process] Taking the SAO short text in the library and information science field as the data source, firstly, a semantic association scheme including the three links of "expansion-reconstruction-denoising" was designed, that is, the semantic information of SAO was extended through semantic expansion and SAO reconstruction, and the noise interference problem after expansion was solved through semantic denoising; then the BERT model was used to train the SAO short text after semantic association; finally, automatic classification was realized in the classification part. [Result/conclusion] After comparing different association values, learning rates and classifiers, the experimental results show that when the association value is 10 and the learning rate is 4e-5, the SAO short text classification effect is optimal, with an average F1 value of 0.8522. Compared with SVM, LSTM and pure BERT, the F1 value is increased by 0.1031, 0.1538 and 0.1405 respectively.

Keywords: SAO; short text classification; semantic association; BERT

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