Postprint: Multi-Feature Fusion for Keyword Semantic Function Recognition

Authors: Zhang Guobiao, Pengcheng Li, Lu Wei, Cheng Qikai

Date: 2023-04-01T00:00:00+00:00

Abstract

[Purpose/Significance] Functional identification of keywords, which serve as lexical items or terms capable of revealing the themes and core content of academic texts, can provide underlying indexing support for the rapid and precise retrieval of knowledge and literature. [Method/Process] To address the limitation of existing research in keyword context modeling that mostly focuses on symbolic semantic representation at the textual level, and through in-depth exploration of literature composition patterns, this paper proposes a lexical function identification model based on multi-feature fusion. The model, while adopting BERT to capture keyword context dependency features, fuses the positional information of keywords in the keyword list and the full text, as well as prior knowledge of lexical functions, and then employs an attention mechanism and feed-forward neural network to perform problem-method semantic function discrimination for keywords. [Results/Conclusion] Experimental results demonstrate that both keyword positional information and prior knowledge can effectively enhance the performance of keyword semantic function identification, with prior knowledge contributing substantially to the improvement in identification effectiveness.

Full Text

Preamble

Research on Keyword Semantic Function Recognition Based on Multi-Feature Fusion

Zhang Guobiao[1],[2], Li Pengcheng[1],[2], Lu Wei[1],[2], Cheng Qikai[1],[2]
[1]School of Information Management, Wuhan University, Wuhan 430072
[2]Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072

Abstract: [Purpose/Significance] Keywords, as vocabulary or terms that reveal the subject and core content of academic texts, can provide underlying

index support for rapid and precise knowledge and document retrieval when their functions are identified. [Method/Process] Existing studies on keyword context modeling are mostly limited to symbolic semantic representation at the textual level. Based on an in-depth analysis of academic writing conventions, this paper proposes a lexical function recognition model based on multi-feature fusion. The model captures keyword context-dependent features using BERT while integrating position information of keywords in both the keyword list and full text, along with prior knowledge of lexical functions. Attention mechanism and feed-forward neural networks are then employed to discriminate the semantic functions of keywords as either research problems or methods. [Result/Conclusion] Experimental results demonstrate that both keyword position information and prior knowledge can effectively improve keyword semantic function recognition performance, with prior knowledge contributing significantly to the improvement.

Keywords: lexical function recognition; academic text; keyword; BERT; multi-feature fusion
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2021.09.010

Keywords serve as functional vocabulary that maps the thematic content of academic literature, providing multi-level semantic tags for information retrieval, knowledge organization, and large-scale text computation. However, the emphasis on conciseness sacrifices substantial contextual information, resulting in ambiguous semantic functions and unclear usage intentions that make interpretation difficult when detached from the original text. Compared to other retrieval methods, queries based on keywords typically require more secondary processing for information filtering and screening [1]. For instance, readers seeking technical details and algorithmic improvements for “BM25” may retrieve numerous documents about applying BM25 to specific problems instead. Therefore, identifying the semantic functions of keywords in academic literature can provide foundational support for targeted rapid knowledge indexing, holding significant theoretical and practical value for precise knowledge retrieval and structured knowledge representation.

Keyword semantic function recognition requires understanding contextual semantics while fully exploiting underlying writing conventions. Existing research on lexical context modeling is largely confined to symbolic semantics at the textual level, neglecting important information such as keyword position and text structure—features that can reveal functional roles to some extent. According to academic writing norms, keywords with different functions should exhibit distinct probability distributions throughout a paper. For example, a research problem like “image classification” would likely be emphasized in the introduction or related work sections, while a research method like “support vector machine” would tend to appear more frequently in methods or experimental sections. W. Lu et al. [3] found through statistical analysis that the ordering of keywords in keyword lists also follows certain patterns: keywords

describing problems or methods are typically positioned at the front of the list, a phenomenon particularly prominent in Chinese journal articles. Furthermore, identical keywords often have established functional tendencies across different disciplines or fields. For instance, "support vector machine" appears as a research problem in machine learning (e.g., improving classifier accuracy and recall) but more frequently serves as a research method in image recognition. To effectively represent and utilize these characteristics, this paper designs representation methods for different features and constructs a multi-feature fusion model for keyword semantic function recognition, integrating keyword position information and prior knowledge to achieve semantic function identification based on comprehensive context feature capture.

2 Related Work Overview

Keyword semantic function recognition falls within the domain of academic text lexical function research, which remains in its preliminary exploration stage. Early lexical function studies focused on Lexical Functional Grammar (LFG) [4], employing statistical natural language processing methods to analyze subject-predicate-object roles at the syntactic level. Widely applied models included Conditional Random Fields (CRF) and Hidden Markov Models (HMM). T. Moon et al. [5] designed an HMM model utilizing boundary conditions by analyzing differences between text content and function words. Sun et al. [6] first performed lexical annotation using a dictionary, then applied CRF for iterative optimization. However, grammatical functions only reflect syntactic relationships between words, failing to capture true semantic meaning. T. Kondo et al. [7] semantically categorized lexical functions into four classes—"domain," "problem," "method," and "other"—for dynamic analysis of technical paths and hotspot evolution in specific fields. Subsequently, H. Nanba et al. [8] expanded the scope to patent and scientific literature abstracts for lexical semantic function recognition. S. Gupta et al. [9] employed syntactic templates and resampling on ACL datasets, achieving significant improvement in domain, problem, and technique categories, though still not reaching practical levels. C. T. Tsai et al. [10] divided lexical functions into technical and application categories, achieving better performance through multi-feature combination and resampling. Cheng et al. [11] further defined concepts of academic literature lexical functions, designing 27 features including lexical, syntactic, and chunk features, and built a CRF-based recognition model for research problems and methods, validated on the GUPTA dataset. K. Heffernan et al. [12] employed multiple machine learning methods including SVM to classify vocabulary as problem or method in academic texts. Lu et al. [13] used BERT and LSTM methods to construct classification models for keyword semantic functions.

Overall, researchers have explored semantic function recognition for titles, abstracts, and keywords in academic texts, achieving some results while revealing limitations. Since lexical semantic function refers to the semantic role vocabulary plays in text [11], existing methods model only contextual information.

However, relying solely on context-dependent features cannot fully represent lexical semantic information; position information in documents and usage conventions in knowledge inheritance also reveal corresponding semantic functions. For example, problem-related vocabulary typically appears in introductions or literature review sections, while method-related vocabulary tends to appear more frequently in methods and experimental sections. Therefore, this paper proposes a keyword semantic function recognition model that combines keyword position information and prior knowledge.

3 Keyword Multi-Feature Representation and Fusion

3.1 Keyword Multi-Feature Representation

3.1.1 Context-Dependent Features

Academic paper titles reveal research topics and innovations, while abstracts concisely state research purposes, methods, results, and conclusions. Recent trends require structured abstracts describing main content. Therefore, title and abstract information can capture keyword context-dependent features. When mentioning problems and methods, authors employ habitual expressions such as "基于/XX 的/…" ("based on XX…"), "…采用/XX 方法…" ("…using XX method…"), and "…是/XX 问题/…" ("…is a XX problem…"), which constitute keyword context features. This paper employs the BERT model, currently the most effective for text processing, to represent keyword context-dependent features.

3.1.2 Position Features

Keywords of different semantic functions should exhibit distinct expression details. Unlike complex open-domain texts, academic texts feature rigorous logical structure and standardized hierarchy, following the general scientific research process from problem introduction to method description to results discussion [11]. A complete research paper typically comprises five sections: introduction, related work, methods, experiments, and conclusion, each serving different structural functions. For instance, the introduction presents background and problems, while related work systematically surveys relevant literature. Consequently, keywords describing research problems frequently appear in introduction and related work sections, while method-describing keywords tend to appear in methods and experimental sections. This paper vectorizes keyword position and frequency information by counting occurrences across sections. Additionally, for keyword list position features, one-hot encoding represents ordering information, setting the position index to 1 and others to 0.

3.1.3 Prior Knowledge Features

Identical keywords often have established functional tendencies across disciplines. For example, "support vector machine" appears as a research problem in machine learning (improving classifier metrics) but more likely as a method in image recognition. This paper represents prior knowledge probabilistically by counting how often a keyword is labeled as problem, method, or other in the domain dataset, generating a 3-dimensional feature vector.

3.2 Keyword Multi-Feature Fusion

Multi-feature fusion combines multiple features into a new vector for lexical function recognition. While simple concatenation fails to consider differences between features, attention mechanisms can distinguish feature importance by assigning weights. The attention function maps a query to key-value pairs through three steps: 1) computing similarity between query and each key for weights; 2) normalizing weights with Softmax; 3) weighted summation of values. The process is shown in formulas (1)-(3):

$$L(H) = tanh(wH + b)$$

$$\alpha = softmax(L(H)) = \frac{exp(L(H))}{\sum_i exp(L(H))_i}$$

$$F = H \cdot \alpha$$

where $L(H)$ represents feature weights, w denotes weight coefficients, b is bias, $tanh$ is the activation function, α represents normalized feature weights, and i indexes vector H dimensions.

The attention mechanism applies weighted transformation to four keyword features, highlighting important features' contributions to improve classification accuracy. As shown in Figure 1 [Figure 1: see original paper], input data X yields four features: f_1 (context-dependent), f_2 (full-text position), f_3 (keyword list position), and f_4 (prior knowledge). The concatenated vector $H = [f_1, f_2, f_3, f_4]$ undergoes $tanh$ weighting and Softmax normalization to obtain attention probabilities α, which are then dot-multiplied with H to produce fused features F.

4 Keyword Semantic Function Recognition Model Construction

As an information extraction task, keyword semantic function recognition can be transformed into text classification. Leveraging deep learning's superior performance, this paper builds a semantic function recognition model using deep learning-based labeling strategies. After fusing context-dependent, position, and prior knowledge features, the model employs deep neural networks for keyword semantic function discrimination, using BERT as the context representation module to optimize nonlinear fitting capability.

4.1 Lexical Function Recognition Model Prototype Selection

Deep learning models typically use pre-trained word embeddings like Word2Vec [14] for vectorization. However, static embeddings lack sentence-level representation and cannot handle polysemy dynamically (e.g., "apple" as fruit vs. company). To address these limitations, Google AI proposed BERT [16] in 2018,

a Transformer [15]-based pre-training method. BERT enhances generalization through character-level, word-level, and sentence-level feature mining. As shown in Figure 2 [Figure 2: see original paper], BERT's representation comprises token embedding, segment embedding, and position embedding. Unlike traditional models, BERT uses deep bidirectional encoding, enabling context-aware dynamic word sense disambiguation by considering both left and right context.

4.2 Multi-Feature Fusion Keyword Semantic Function Recognition Model

To accurately understand keyword semantic functions, this paper fuses context-dependent, position, and prior knowledge features based on BERT. After vectorizing multiple features, concatenation and attention mechanism allocate weights, followed by a Softmax classifier for final classification. The model architecture comprises input, feature representation, feature fusion, and detection layers, as shown in Figure 3 [Figure 3: see original paper].

The **input layer** extracts and preprocesses raw data. For context features, titles, abstracts, and keywords are concatenated and tokenized. For full-text position information, sections are classified into five functional blocks (introduction, related work, methods, experiments, conclusion) using methods from [17][18], with keyword occurrence frequencies counted. Keyword list positions are obtained through matching. Prior knowledge is derived from frequency statistics of function labels in the training dataset.

The **feature representation layer** vectorizes preprocessed data. Context-dependent features use Google pre-trained Chinese BERT, outputting 768-dimensional vectors. Full-text position frequencies are normalized to [0,1] as 5-dimensional vectors. Keyword list positions are represented as 5-dimensional one-hot vectors (based on average keyword count of 5). Prior knowledge frequencies are normalized to 3-dimensional vectors.

The **feature fusion layer** employs the attention mechanism from Section 3.2 for weighted fusion. The **output layer** uses two fully connected layers with Softmax classification.

5 Keyword Semantic Function Recognition Experiments

5.1 Data Annotation

No standard corpus exists for academic keyword semantic function recognition. This study uses a self-built dataset from 100,025 research papers (2009-2018) in Chinese computer science journals (*Computer Engineering*, *Computer Science*, *Journal of Computer*, *Pattern Recognition and Artificial Intelligence*). Keywords were labeled using title/abstract patterns and manual annotation. A simple classification scheme categorized keywords as research problem, research method, or other. Pattern-based labeling used templates (e.g., "基于 XX 的 XX" - "XX-based XX") with manual review. Irregular cases were manually annotated by

two PhD and two master's students in information science through two rounds: individual annotation followed by voting on uncertain cases. The final dataset contains 310,214 labeled keywords (102,278 methods, 102,504 problems, 105,432 others), split 8:1:1 for training, validation, and test sets (Table 1).

5.2 Experimental Setup

Experiments ran on Ubuntu 16.04 with Python 3.6 and TensorFlow. Four keyword features were extracted: f_1 using Chinese BERT pre-trained model on concatenated title/abstract/keyword sequences (max length 512, 768-dimensional output); f_2 using abstract as full-text proxy (5 sentences, 5-dimensional normalized frequency vector); f_3 as 4-dimensional one-hot keyword list position; f_4 as 3-dimensional normalized prior knowledge vector. Features were concatenated and input to an Attention-based dual fully-connected classifier. Training used exponential learning rate decay (5% per 500 steps), Dropout, EarlyStopping, and Talos hyperparameter optimization. Final parameters are shown in Table 2 .

5.3 Results and Analysis

Evaluation uses Accuracy, Precision, Recall, and F1-measure. Experiments aim to: (1) validate attention fusion effectiveness; (2) compare feature combinations; (3) analyze performance differences across semantic function types.

Attention Effectiveness: Table 3 shows concatenation+Attention outperforms simple concatenation across all metrics, demonstrating Attention's ability to distinguish feature importance and highlight critical information.

Feature Combination Impact: Table 4 compares models with different feature combinations. All features improve F1 scores, with the full-feature model achieving optimal performance (Recall: 0.973, Precision: 0.973, F1: 0.973, Accuracy: 0.978). Prior knowledge contributes most significantly, followed by position features which improve Precision more than Recall.

Semantic Function Type Comparison: Table 5 shows method keywords achieve highest performance (Precision: 0.98, Recall: 0.98, F1: 0.98) due to more standardized expressions, while problem keywords show more variability and complexity, yielding slightly lower metrics.

Conclusion

This paper proposes a multi-feature fusion model for keyword semantic function recognition, integrating context-dependent, position, and prior knowledge features with deep learning. Experiments on a computer science journal dataset demonstrate average Precision of 0.973, outperforming partial-feature methods. Future work will expand the dataset, explore more comprehensive semantic function types, and apply lexical functions to literature recommendation, keyword extraction, automatic summarization, and knowledge graph construction.

References

[1] Yang Tao. Analysis of Chinese intelligent search engines [J]. Library and Information Service, 2002, 56(1): 62-65.
[2] Wan Hualin, CHOWDHURY M U. Image semantic classification based on support vector machine [J]. Journal of Software, 2003, 14(11): 1891-1899.
[3] LU W, LI X, LIU Z F, et al. How do author-selected keywords function semantically in scientific manuscripts [J]. Knowledge Organization, 2019, 46(6): 403-418.
[4] DALRYMPLE M, KAPLAN R M, MAXWELL J T, et al. Formal issues in lexical-functional grammar [M]. Stanford: CSLI Publications, 1995.
[5] MOON T, ERK K, BALDRIDGE J. Crouching dirichlet, hidden markov model: unsupervised POS tagging with context local tag generation [C]//Proceeding of the 2010 conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2011: 196-206.
[6] Sun Jing, Li Junhui, Zhou Guodong. Unsupervised Chinese POS tagging based on conditional random fields [J]. Computer Applications and Software, 2011, 28(4): 21-23.
[7] KONDO T, NANBA H, TAKEZAWA T, et al. Technical trend analysis by analyzing research papers' titles [M]. Berlin: Springer, 2009.
[8] NANBA H, KONDO T, TAKEZAWA T. Automatic creation of a technical trend map from research papers and patents [C]//Proceedings of the 3rd international workshop on patent information retrieval. New York: Association for Computing Machinery, 2010: 11-18.
[9] GUPTA S, MANNING C. Analyzing the dynamics of research by extracting key aspects of scientific papers [C]//Proceedings of 5th international joint conference on natural language processing. Stroudsburg: Association for Computational Linguistics, 2011: 1-9.
[10] TSAI C T, KUNDU G, DAN R. Concept-based analysis of scientific literature [C]//Proceeding of the 22nd acm international conference on conference on information & knowledge management. New York: Association for Computing Machinery, 2013: 1733-1738.
[11] Cheng Qikai, Li Xin. Automatic recognition of lexical semantic functions for semantic publishing [J]. Digital Library Forum, 2017, 159(8): 24-31.
[12] HEFFERNAN K, TEUFEL S. Identifying problems and solutions in scientific text [J]. Scientometrics, 2018, 116(1): 1-16.
[13] Lu Wei, Li Pengcheng, Zhang Guobiao, et al. Lexical function recognition in academic text: automatic keyword classification based on BERT vector representation [J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(12): 1320-1329.
[14] Tang Ming, Zhu Lei, Zou Xianchun. A document vector representation based on Word2Vec [J]. Computer Science, 2016, 43(6): 214-217, 269.
[15] VASWANI A, SHAZEER N, PARMAR N. Attention is all you need [C]//Proceedings of the 31st international conference on neural information processing systems. New York: Curran Associates Inc, 2017: 6000-6010.

[16] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Proceeding of the 2019 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.
[17] Wang Jiamin, Lu Wei, Liu Jiawei, et al. Multi-level fusion for academic text structure function recognition [J]. Library and Information Service, 2019, 63(13): 95-104.
[18] Huang Yong, Lu Wei, Cheng Qikai, et al. Structure function recognition of academic text: paragraph-based recognition [J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(5): 530-538.

Author Contributions

Zhang Guobiao: paper writing, data annotation, experimental analysis
Li Pengcheng: data annotation, experimental design
Lu Wei: research framework, paper revision
Cheng Qikai: experimental design, paper revision

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