Dictionary Construction for Semantic Features of English Scientific Paper Abstracts (Postprint)

Authors: Song Donghuan, Li Chenying, Liu Ziyu, Han Mingjie

Date: 2023-04-01T16:15:51+00:00

Abstract

[Purpose/Significance] Paper abstracts constitute important indexing objects for information organization. Structurally indexing paper abstracts facilitates scientific communication, knowledge discovery, and intelligence analysis. Achieving accurate and rapid automatic indexing of existing unstructured abstracts represents a practical problem requiring urgent resolution. [Method/Process] Assuming intrinsic consistency across different abstract categories, research on structured abstracts can provide methodological and technical references for automatic indexing of unstructured abstracts. Accordingly, based on the National Library of Medicine's structural element label terminology set and label classification mapping relationships, a structural element BOMRC system and an identification and standardized indexing method for structured abstracts are proposed. Subsequently, research samples are selected and text mining methods are employed to conduct quantitative statistical analysis of multiple indicators—including word frequency and TFIDF values—for words, verbs, three-word chunks, four-word chunks, and other lexical units in the sample corpus, constructing a semantic feature dictionary capable of structural element identification. Finally, an unstructured abstract test set is utilized to validate the effectiveness of the semantic feature dictionary. [Results/Conclusion] The results demonstrate that the semantic feature dictionary method can effectively identify various elements of unstructured abstracts and can be applied to optimize automatic identification models centered on machine learning methods.

Full Text

Preamble

Volume 64, Issue 6, March 2020
ChinaXiv Cooperative Journal

Construction of a Semantic Feature Dictionary for English Scientific Paper Abstracts

Song Donghuan[1],[2], Li Chenying[1], Liu Ziyu[1], Han Mingjie[1]
[1]China Agricultural University Library, Beijing 100193
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract:
[Purpose/Significance] Abstracts serve as crucial indexing objects for information organization. Structuring abstracts according to specific frameworks facilitates scientific communication, knowledge discovery, and intelligence analysis. Automatically indexing existing unstructured abstracts with precision and speed represents an urgent practical challenge. [Method/Process] This study assumes that different abstract categories possess intrinsic consistency, meaning research on structured abstracts can provide methodological and technical references for automatic indexing of unstructured abstracts. Accordingly, based on the U.S. National Library of Medicine's structured element labeling terminology set and label classification mapping relationships, we propose the BOMRC (Background-Objective-Method-Result-Conclusion) element system and a method for identifying and normalizing structured abstracts. We then selected research samples and employed text mining methods to conduct quantitative statistical analysis of multiple indicators—including word frequency and TF-IDF values—for words, verbs, three-word chunks, and four-word chunks in the sample corpus, constructing a semantic feature dictionary capable of identifying structural elements. Finally, we tested the dictionary's effectiveness using an unstructured abstract test set. [Result/Conclusion] Results demonstrate that the semantic feature dictionary method can effectively identify various elements in unstructured abstracts and can be used to optimize automatic recognition models centered on machine learning methods.

Keywords: scientific paper; paper abstract; structural element; semantic feature; feature dictionary
Classification Number: G254
DOI: 10.13266/j.issn.0252-3116.2020.06.013

Since the 21st century, research papers have grown rapidly, making information overload a practical problem troubling academia. Enabling users to quickly and accurately discover needed papers has become a research direction for numerous information service institutions, including publishing and library and information science communities. Scientific paper abstracts possess strong purposiveness and structural functions, serving as concise summaries of paper content and as important bases for readers to retrieve and filter papers. Moreover, abstracts are vital indexing objects for information organization, receiving considerable attention from indexing databases, while their in-depth text mining and automatic indexing have also attracted interest from library and information research and computer technology application studies.

Currently, scientific journal paper abstracts fall into two major categories: struc-

tured and unstructured abstracts [1]. Compared with unstructured abstracts, which suffer from inconsistent formatting, unclear hierarchy, incomplete content, and difficulties in text mining, structured abstracts demonstrate prominent advantages in completeness, clarity, information volume, and suitability for shallow reading in mobile environments. Consequently, they are being adopted by an increasing number of journals. According to a 2018 survey by our research group covering 1,900 journals under ESI subject categories, we systematically sampled 20% (380 journals) ranked by impact factor and found that 188 journals (49.47%) had adopted structured abstracts. However, journals using structured abstracts remain a minority, making the classification and indexing of research purposes, primary methods, important results, and conclusions in unstructured abstracts—along with deep mining of key content—an important task facing information organization and service providers.

This study posits that different abstract categories possess intrinsic consistency, meaning structured and unstructured abstracts show high consistency in writing conventions, habitual expressions, and writing objectives. Therefore, research on structured abstracts can provide methodological and technical references for automatic indexing of unstructured abstracts. To this end, we began with existing journals that have adopted structured abstracts, summarized 157 types of structured abstract element labels and 299 label combination patterns, and proposed a mappable "Background-Objective-Method-Result-Conclusion (BOMRC)" five-element system. We then studied the lexical attribute characteristics of current structured abstracts, using text mining and quantitative analysis to construct a semantic feature dictionary. Finally, we developed an abstract indexing model based on the feature dictionary and tested it on manually annotated unstructured abstract corpora. The value of this study lies in providing a feature dictionary as an information organization tool foundation for standardized indexing of structured abstracts and rapid identification and indexing of structural elements in unstructured abstracts. This dictionary can also optimize existing automatic indexing models and explain automatic indexing results, significantly improving the accuracy and interpretability of unstructured abstract automatic indexing and offering solutions for indexing millions of scientific paper abstracts.

2 Related Research

Through comprehensive review of research papers on abstract content and semantics, along with backward and forward citation tracking, we collected 1,526 relevant English papers and 613 Chinese papers. These studies were primarily published in computer science, journal editing, and applied linguistics journals, focusing on two aspects:

2.1 Research on Abstract Elements

2.1.1 Linguistic Features of Elements Linguistic features broadly include tense, voice, word order, word count, and vocabulary. Tense and voice stud-

ies focus on analyzing new trends in abstract writing, while word order, word count, and vocabulary analyses derive characteristic patterns for abstract element writing by analyzing certain abstract samples. For example, Cao Yan et al. [2] used the IMRD (Introduction-Method-Results-Discussion) four-element model as the analysis object, employing Range vocabulary analysis software to tag vocabulary under each element and discovered that each element contains some preferential lexical chunks. R.A. Day et al. [3] investigated tense usage frequency across elements and found that Method and Results sections show similar tense application, with past tense being used more frequently. Qian Duoxiu et al. [4] conducted comparative studies on abstract elements and found that IMRD elements show a trend toward using simple present tense.

2.1.2 Pattern Features of Elements This research primarily focuses on the number and combination of elements. Scholars represented by N. Graetz [5] first proposed a four-element model, summarizing a universal "Problem-Method-Results-Conclusions" pattern. Subsequently, J.M. Swales [6] questioned the reliability and scientific validity of Graetz's research data, arguing that abstract element patterns should correspond one-to-one with paper element patterns and advocating that abstracts should consist of IMRD four elements. Meanwhile, F. Tseng [7], Li Tao [8], and Zhou Zhichao [9] proposed several other four-element variants based on the IMRD pattern, such as the BMRC (Background-Method-Result-Conclusion) model. However, some scholars found that to ensure abstract completeness, background introduction should be added. Therefore, T. Dahl [10] proposed a five-element "Background-Purpose-Methodology-Result-Comments on results" model based on Swales' framework. Additionally, R.B. Haynes [11] proposed a seven-element "Objective-Design-Setting-Patients or participants-Interventions-Measures and Results-Conclusion" model for medical papers as early as 1987. Currently, many medical journals provide multiple structured abstract writing requirements based on article types, with JAMA Surgery and Physiotherapy requiring up to eight elements in structured abstracts.

2.2 Research on Abstract Semantic Features

American semantic experts L.F. Don and A.P. Nilsen [12] proposed that semantic features include five categories: grammatical-semantic features, intrinsic semantic features, predicative semantic features, adverbial semantic features, and perceptual semantic features. When analyzing semantic features of vocabulary or other entities, researchers often use "[+/- semantic attribute]" to represent corresponding semantic features. The second and fifth categories belong to lexical-level semantic features, where element categories can serve as semantic feature attributes, such as [+background][-objective][-method][-result][-conclusion]. The remaining categories belong to grammatical-level semantic feature analysis, which cannot be analyzed in isolation from sentences, as individual words cannot express any semantics.

Research on using semantic feature technology for abstract element identification includes: single-feature-based semantic recognition technology research and comprehensive-feature-based semantic recognition technology research. Single-feature research refers to using only one feature such as word frequency, word order, or tense for abstract element semantic identification. In 2002, L.E. Anthony [13] first constructed an automatic abstract identification model, initially using small abstract datasets to extract continuous word clusters of one to five words and applying Naive Bayes algorithms for learning to identify structural element content. S.N. Kim and L. Martinez [14] found that when using word order for structural element identification, Conditional Random Fields algorithms perform better than Naive Bayes and Support Vector Machine algorithms, with accuracy generally above 90%. Comprehensive-feature semantic recognition technology relies more on human subjectivity compared to single-feature technology, requiring manual selection of features for analysis. For example, V.D. Feltrim et al. [15] divided abstracts into several sentence groups and conducted structural element identification research based on sentence position. J. Silva et al. [16] and Y.K. Meena et al. [17] also constructed different types of element identification models using sentence features. Y. Guo et al. [18] used syntactic analysis tools to compare lexical and contextual features, finding that lexical features have the best predictive effect, while voice and element order identification perform worst. Shen Si et al. [19] used characters as basic semantic units and built an automatic identification model for journal paper abstract structural functions based on LSTM-CRF deep learning methods, though their structural element selection did not consider disciplinary differences.

In summary, we found that: lexical chunks have unique characteristics, and lexical feature prediction performs better than voice and element order identification; vocabulary attributes can be regarded as manifestations of semantic features, while intrinsic semantic features mainly focus on basic conceptual and logical semantic features; abstract element identification research has focused on semantic features such as word frequency, tense, voice, and position. However, no research has been reported on constructing semantic feature dictionaries from the perspective of vocabulary attributes to complete abstract content tagging. Therefore, to address issues of strong subjective dependence, feature sparsity, and limited interpretability in previous studies, this study attempts to construct a semantic feature dictionary guided by quantitative analysis and vocabulary attributes, establishing an information organization foundation for abstract element identification.

This study uses structured abstract data from English scientific journals as samples to deeply mine structural elements and textual features of structured abstracts through specific research on three questions: How to determine identification and indexing methods for structured abstracts? Do representative feature vocabulary with semantic identification functions exist in structural elements? Can feature words identify structural elements in unstructured abstract sentences, and how effective is this identification? We first analyzed and summarized structured abstract features, finding that over 35% of structured

abstracts adopt BMRC or OMRC label combination patterns. We also discovered that the U.S. National Library of Medicine's structured abstract label terminology set maps labels according to BOMRC classification relationships. Therefore, we proposed using the BOMRC five-element model for structured abstract identification and indexing. Then, by calculating content feature words in structured abstract sentences uniformly mapped to the BOMRC model, we extracted feature word candidate sets under different structural elements. Using a structured abstract test set, we completed correction and refinement of the feature word candidate set, constructing a semantic feature dictionary applicable for BOMRC structural element tagging. Finally, we conducted validity testing using the semantic feature dictionary. The specific research content and design are shown in Figure 1 [Figure 1: see original paper].

3 Research Methods and Data Preparation

3.1 Research Objectives and Design

To improve intelligent indexing levels in information organization and provide references for computer-based normalized indexing of structured abstracts, structured indexing of unstructured abstracts, and information extraction in intelligence analysis, this study (research design shown in Figure 1 [Figure 1: see original paper]).

3.2 Data Preparation

3.2.1 Preparation of Basic Journal Paper Data This study used Clarivate Analytics' Web of Science Core Collection database to collect TOP10% highly cited papers and their citing papers from nine fields under the Chinese Academy of Engineering's "Global Engineering Frontier Research Project," totaling 284,525 records (2012-2017). After excluding non-Article document types and extracting the latest issue paper data for each journal, we obtained 16,900 (13,046 after deduplication) brief information records for 10,043 (7,218 after deduplication) journals, as shown in Table 1 .

3.2.2 Screening of Structured Abstract Papers

(1) Using the 3,032 structured abstract labels provided by the U.S. National Library of Medicine as the label terminology set [20], combined with special marker character features following structured abstract element labels collected during this study, we employed simple pattern matching and forward consistent matching methods. Based on characteristics such as label position and quantity in abstracts, we screened 13,046 abstract data records, selecting 1,583 papers using structured abstracts, as shown in Figure 2 [Figure 2: see original paper].

(2) Manual verification of abstracts slated for removal revealed that some structured abstracts contained issues such as label spelling errors, special

label formats, or labels not included in the NLM terminology set. After manually supplementing 11 inaccurately labeled papers, we marked 1,594 relatively recent papers from 1,213 journals as structured abstract research samples. Examples of structured abstract identification and indexing results are shown in Figure 3 [Figure 3: see original paper].

(3) We selected all 13,781 paper data records from 1,213 journals in the dataset for another structured abstract screening, which yielded 5,021 papers with unstructured abstracts (serving as the unstructured abstract detection set) and 8,760 papers with structured abstracts. Excluding the 1,594 previously studied structured abstract samples, the remaining 7,166 papers served as the structured abstract test set for subsequent research. We also supplemented and mapped labels not included in the terminology set, adding 30 labels.

4 Semantic Feature Attribute Classification Research

4.1 Distribution of Structural Element Labels

Statistics on the 157 element labels used in 1,594 papers' structured abstracts revealed that all labels appeared 6,582 times, with each label appearing in an average of 42 abstracts. Conclusion appeared most frequently, followed by Result, Method, Background, and Objective, as shown in Figure 4 [Figure 4: see original paper].

4.2 Distribution of Structural Element Label Combination Patterns

Statistical analysis of label combination patterns revealed 299 types. The most frequent pattern was "Background + Method + Result + Conclusion" (BMRC), accounting for over one-quarter of cases. The second most common patterns were "Objective + Method + Result + Conclusion" (OMRC) and "Background + Result + Conclusion" (BRC), as shown in Figure 5 [Figure 5: see original paper].

This result abandons the analysis method of examining single-element proportions [9] and instead provides detailed statistics on overall combination pattern proportions, which better facilitates mining of sequential relationships among elements. We also found that the five-element combination of Conclusion + Result + Method + Background + Objective is most universal. Through investigation of numerous journal submission guidelines and definitions of research elements in literature, we summarized concepts contained in the BOMRC five elements (see Figure 6 [Figure 6: see original paper]). According to the defined element definitions, we found that all structured abstract element labels can be mapped under these five elements. This not only ensures completeness of abstract content identification but also distinguishes core content, making vocabulary attribute-based identification of BOMRC five-element content highly significant.

5 Construction of Semantic Feature Dictionary

This study aims to obtain feature vocabulary capable of structured identification and indexing of unstructured abstract text content. Since identifying features in unstructured abstract text content is relatively simple when using sentences as units, we treated all previously obtained structured abstract texts as documents at the sentence level. Combined with mapping relationships summarized by the U.S. National Library of Medicine, we categorized extracted elements into BOMRC five-element classifications to screen feature vocabulary capable of identifying sentence structural elements. Additionally, the TF-IDF method is currently the most commonly used text feature weighting method, effectively combining local and global weights to identify words that appear frequently in one document but rarely in other documents across the collection. Therefore, this study conducted feature word screening through the following processes and calculation methods.

5.1 Construction of Semantic Feature Dictionary Candidate Set

Based on the frequency and TF-IDF value proportion of each vocabulary item in the BOMRC five-element document collection, we manually observed numerical values and interval distributions of various indicators. We discovered that screening feature words with potential structural identification functions according to the following thresholds (see Table 2) was effective:

- **Words:** Frequency ≥ 5 and TF-IDF proportion $\geq 50\%$ (excluding cases where TF-IDF $\geq 40\%$ in other four elements)
- **Verbs:** Frequency ≥ 3 and TF-IDF proportion $\geq 50\%$ (excluding cases where TF-IDF $\geq 40\%$ in other four elements)
- **Three-word chunks:** Frequency ≥ 2 and TF-IDF proportion $\geq 50\%$ (excluding cases where TF-IDF $\geq 40\%$ in other four elements)
- **Four-word chunks:** Frequency ≥ 2 and TF-IDF proportion $\geq 50\%$ (excluding cases where TF-IDF $\geq 40\%$ in other four elements)

Through these initial screening criteria, we selected 15,526 feature words from 481,877 vocabulary items to form a feature word candidate set. However, due to attribute inconsistencies in feature words with inclusion relationships, we analyzed five situations: word-three-word, word-four-word, verb-three-word, verb-four-word, and three-word-four-word combinations, removing 451 feature words with inconsistent attributes (e.g., "suggest" had Conclusion attribute while "recent studies suggest that" had Background attribute). We retained 15,075 feature words (see Table 3).

5.2 Correction of Semantic Feature Dictionary Candidate Set

Feature word annotation accuracy is the core factor of dictionary effectiveness. Therefore, improving the average annotation accuracy of vocabulary in the candidate set is the focus. Using 7,166 structured abstract test set papers as corpus for correction and refinement, we performed sentence segmentation and removed

content unrelated to abstracts, such as copyright information, links, and email addresses, obtaining 80,346 sentences. We statistically analyzed feature words appearing in each sentence and their annotation label accuracy (correctly labeled as 1, incorrectly labeled as -1, unlabeled as 0). We also divided the overall annotation accuracy proportion of each feature word into intervals and verified the results obtained using the feature dictionary candidate set against original labels in structured abstract sentences. Feature words with annotation accuracy below 50% were removed, retaining 6,447 feature words. Table 4 shows the distribution of four types of feature vocabulary across different accuracy intervals, clearly demonstrating that three-word feature chunks have significantly higher annotation accuracy than other types.

5.3 Refinement of Semantic Feature Dictionary Candidate Set

Since feature word selection primarily relied on word frequency and TF-IDF proportion, the basic characteristics of corrected candidate set vocabulary could be analyzed from two aspects: feature word frequency proportion and TF-IDF ranking interval proportion. The calculation process for specific indicators is as follows:

- **Frequency proportion:** Count frequencies of all vocabulary in each element; Calculate the proportion of vocabulary frequency under each element in the total corpus frequency for the same word; Statistically analyze the frequency proportion intervals of feature vocabulary.
- **TF-IDF ranking interval:** Calculate TF-IDF values for all vocabulary in each element; Rank and number all vocabulary under the same element by TF-IDF value from low to high; Normalize all numbers and statistically analyze TF-IDF ranking intervals (ranking interval = vocabulary ranking value / total vocabulary frequency under the element).

Analysis of corrected feature vocabulary frequency proportion and TF-IDF ranking interval in BOMRC five elements (see Figure 8 [Figure 8: see original paper], where 1: Background, 2: Objective, 3: Method, 4: Result, 5: Conclusion) revealed that frequency proportions of feature words, feature verbs, three-word feature chunks, and four-word feature chunks mainly concentrated in 65%-100%, 60%-100%, 60%-100%, and 60%-100% respectively. TF-IDF ranking intervals concentrated in top 30%, top 45%, top 10%, and top 5% respectively. Therefore, we used comprehensive indicators including word frequency, frequency proportion, TF-IDF value, and TF-IDF ranking interval as supplementary criteria for feature words. Due to increased sentence numbers in the test set, word frequency required adjustment. Combined with manual observation, we ultimately determined new criteria: word frequency ≥ 100, other vocabulary types frequency ≥ 5. Using the structured abstract test set to calculate vocabulary and metrics, we supplemented 5,542 feature words, bringing the total dictionary vocabulary to 11,989. After attribute statistics and removing inconsistent attributes in inclusion relationships, we finalized 11,761 feature words, expanding the refined feature dictionary to 11,761 vocabulary items.

6 Validity Testing of Semantic Feature Dictionary

The identification effectiveness of the semantic feature dictionary requires validation using an unstructured abstract detection set. Therefore, we used the extracted 5,021 unstructured abstract papers as the detection set, obtaining 43,517 sentences through sentence segmentation and removal of irrelevant sentences. The semantic feature dictionary was used to annotate each sentence with corresponding element labels, and identification effectiveness was comprehensively evaluated through sampling combined with manual verification.

6.1 Analysis of Semantic Feature Dictionary Annotation Results

Using the semantic feature dictionary for exact matching, we labeled 29,530 sentences (67.86%) with element tags. A total of 11,761 feature words participated in machine annotation (73.12%). The distribution of matched feature words is shown in Table 5 .

Analysis of labeled sentence numbers (see Table 6) revealed that three-word feature chunks labeled the most sentences, while four-word feature chunks labeled the fewest. Observation of sentence label combinations found that "Method + Result" was the most common combination (2,552 instances), indicating that Method and Result element feature words co-occurred in sentences. This partially resulted from Method sections containing outcome-related terms such as "Outcome" and "Outcome measurement" that characterize clinical trial results. Annotation result examples are shown in Figure 9 [Figure 9: see original paper].

6.2 Manual Annotation and Sampling

After machine annotation, we invited two annotators for result verification. First, we extracted 10 abstracts to demonstrate the manual annotation process. Since annotation was limited to BOMRC five elements, we used the previously mentioned element definitions as annotation standards. After training, both annotators independently annotated the same 20 unstructured abstracts for pilot testing. We discussed inconsistent results and reached consensus. According to these standards, we divided the 5,021 unstructured abstract papers by field, first allocating by paper proportion per field, then sorting journals by ISSN in ascending order and selecting corresponding journals at equal intervals (interval = 5), yielding 4, 4, 4, 2, 4, 6, 26, 40, and 8 journals respectively. For each journal, we selected the first paper abstract sorted by unique metadata identifier as the manual annotation object. Following these criteria, we selected 98 papers for manual annotation.

6.3 Verification of Semantic Feature Dictionary Identification Effectiveness

Comprehensive evaluation of five-element identification effectiveness showed that the feature dictionary performed worst on Objective identification, largely

because Objective and Background content frequently overlap. However, the average F1 value for identifying all five elements was 0.7606, comparable to the 0.7819 average F1 value reported by Wang Lifei et al. in 2017 [21] for an abstract element identification model combining machine learning algorithms with various linguistic features, proving the identification effectiveness of the semantic feature dictionary. The dictionary showed better identification effectiveness for Method and Result elements.

This study employed a traditional dictionary method to identify structural elements in unstructured abstracts, considering both the method's strong accuracy and interpretability, and that our results could serve as rules to improve and enhance the efficiency of existing automatic indexing models built with machine learning algorithms. This study addressed three tasks: determining structured abstract identification and normalization indexing methods, constructing a semantic feature dictionary for identifying sentence structural element categories, and verifying the dictionary's effectiveness in identifying structural element attributes. The constructed semantic feature dictionary fully validated our initial research hypothesis, demonstrating the intrinsic consistency among different abstract types and providing new ideas for identifying other element patterns in unstructured abstracts.

This study makes three contributions: It not only determined structured abstract identification and normalization indexing methods but also enriched the structured abstract element label library and mapping relationships; It identified representative feature vocabulary—including feature words, feature verbs, three-word feature chunks, and four-word feature chunks—with semantic identification functions in structural elements, constructing a semantic feature dictionary containing four vocabulary types; The dictionary's identification effectiveness is comparable to existing automatic abstract element identification models, proving its validity.

However, this study's structured abstract sample size was limited, and the semantic feature dictionary's vocabulary requires further expansion. Due to time constraints, we did not construct sentence templates based on feature word co-occurrence relationships, requiring further research on structural element feature identification through sentence template construction. Additionally, the annotation sample size in validity testing was insufficient, limiting testing effectiveness. Future research should further mine typical sentence patterns with identification functions and explore methods for deep indexing and intelligent indexing of unstructured abstract text content, laying a methodological foundation for effective utilization of scientific journal abstracts.

References

[1] Ertl N. New way of documenting scientific data from medical publications [J]. Kargergazette, 1969, 27(20): 1-3.
[2] Cao Yan, Mou Aipeng. A study on the move characteristics of academic

vocabulary in English abstracts of scientific journals [J]. Foreign Language Research, 2011(3): 46-49.
[3] Day RA, Sakaduski N. Scientific English: a guide for scientists and other professionals [M]. Phoenix, AZ: Oryx, 1998: 109-136.
[4] Qian Duoxiu, Luo Yuan. A corpus-based comparative study on moves in research article abstracts [J]. Journal of University of Science and Technology Beijing (Social Sciences Edition), 2014, 30(2): 12-17.
[5] Graetz N. Teaching EFL students to extract structural information from abstracts [M]. Belgium: ACCO, 1985: 123-135.
[6] Swales JM. Genre analysis: English in academic and research settings [D]. Cambridge: Cambridge University Press, 1990.
[7] Tseng F. Analyses of move structure and verb tense of research article abstracts in applied linguistics [J]. International journal of English linguistics, 2011, 1(2): 27-39.
[8] Li Tao. Research on standardization of English abstracts in scientific papers—taking natural science papers as an example [J]. Journal of Liaoning University of Technology (Social Science Edition), 2018, 20(6): 70-73.
[9] Zhou Zhichao. Analysis of core elements and logical structure of abstracts in Chinese LIS journals [J]. Information Science, 2018, 36(3): 8-12, 32.
[10] Dahl T. Lexical cohesion-based text condensation: an evaluation of automatically produced summaries of research articles by comparison with author-written abstracts [D]. Bergen: University of Bergen, 2000.
[11] Haynes RB. A proposal for more informative abstracts of clinical articles [J]. Annals of internal medicine, 1987, 106(4): 598-604.
[12] Nilsen DLF, Nilsen AP. Semantic theory: a linguistic perspective [M]. Massachusetts: Newbury House Publishers, 1975: 1-138.
[13] Anthony LE. A machine learning system for the automatic identification of text structure, and application to research article abstracts in computer science [D]. Birmingham: Birmingham University, 2002.
[14] Kim SN, Martinez D, Cavedon L, et al. Automatic classification of sentences to support evidence based medicine [J]. BMC bioinformatics, 2011, 12(2): 1-10.
[15] Feltrim VD, Teufel S. Automatic critiquing of novices' scientific writing using argumentative zoning [C]//Proceedings of AAAI spring symposium on exploring attitude and affect in text: theories and applications. 2004, 3: 1-4.
[16] Silva J, Coheur L, Mendes AC, et al. From symbolic to sub-symbolic information in question classification [J]. Artificial intelligence review, 2011, 35(2): 137-154.
[17] Meena YK, Gopalani D. Feature priority based sentence filtering method for extractive automatic text summarization [J]. Procedia computer science, 2015, 48(1): 728-734.
[18] Guo Y, Korhonen A, Liakata IM, et al. Identifying the information structure of scientific abstracts: an investigation of three different schemes [C]//Proceedings of the 2010 workshop on biomedical natural language processing. Association for Computational Linguistics, 2010: 99-107.
[19] Shen Si, Hu Haotian, Ye Wenhao, et al. Research on automatic identifi-

cation of abstract structural functions based on full-character semantics [J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(1): 79-88.
[20] US National Library of Medicine. The NLM label list and category mappings [EB/OL]. [2020-01-02]. https://structuredabstracts.nlm.nih.gov/.
[21] Wang Lifei, Liu Xia. Construction of an automatic identification model for move structure in English academic paper abstracts [J]. Technology Enhanced Foreign Language Education, 2017(2): 45-50, 64.

Author Contributions:
Song Donghuan: Responsible for paper writing and data processing;
Li Chenying: Responsible for research design, paper framework design, writing guidance, and final approval;
Liu Ziyu: Guided some data processing methods and paper revision and improvement;
Han Mingjie: Responsible for some data processing work and paper writing guidance.

[1]China Agricultural University Library, Beijing 100193
Song Donghuan[1],[2], Li Chenying[1], Liu Ziyu[1], Han Mingjie[1]
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190

Abstract: [Purpose/significance] The abstract of scientific papers is a vital indexing object within information organization. Indexing the abstract according to certain structures is conducive not only for scientific communication or knowledge discovery, but also for intelligence analysis. Thus, how to realize auto-index accurately and quickly for existing unstructured abstracts is a crucial problem to be addressed. [Method/process] This study assumed that different categories of abstracts are inherently consistent, that is, the study of structured abstracts can provide a method and technical reference for unstructured abstract auto-indexing. Acting in accordance with this assumption and based on the US National Library of Medicine's structural element labeling terminology, this study accomplished mapping across abstract element classifications and proposed BOMRC system, a normalization indexing method for structured abstracts. Then we collected research sample and used text mining method to analyze multiple features of structured abstract quantitatively and statistically, such as word frequency, TF-IDF value, as dimension of words, verbs, three-word lexical chunks and four-word lexical chunks, which enabled us propose a semantic feature dictionary for structured elements. Finally, we used unstructured abstract to test the validity of the semantic feature dictionary. [Result/conclusion] The results show that the semantic feature dictionary method can effectively identify various structural elements of scientific paper abstract, and it can be used to optimize the automatic recognition model, which may be based on machine learning methods.

Keywords: scientific paper; paper abstract; structural element; semantic feature; feature dictionary