Bibliometric Analysis of Machine Learning in Term Extraction Research (Postprint)

Authors: Qiu Keda, Ma Jianling

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

Abstract

[Purpose/Significance] This study systematically reviews and summarizes research on machine learning-based automatic term extraction, providing a reference for researchers and practitioners in the field.

[Method/Process] Utilizing the analytical tools of CNKI and EndNote, bibliometric methods were employed to conduct a macro-level analysis of annual trends and core institutions related to the topic. Subsequently, a thematic content analysis was performed from three perspectives: extraction techniques and methods, datasets and evaluation, and applications.

[Results/Conclusion] In recent years, term extraction research has made significant progress and constitutes foundational work in domains such as knowledge systems, natural language processing, and information analysis. With the rapid development of natural language processing, extraction technologies have begun to evolve toward deep learning. However, the fundamental theoretical framework of term extraction remains to be improved, particularly regarding evaluation metrics, corpus selection, and effectiveness evaluation methods.

Full Text

A Bibliometric Analysis of Machine Learning in Term Extraction Research

Qiu Keda, Ma Jianling

Lanzhou Documentation and Information Center, Chinese Academy of Sciences, Lanzhou 730000
Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000

Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049

Abstract:
[**Purpose/Significance**] This paper aims to review and summarize relevant research on automatic term extraction based on machine learning, providing a reference for researchers in the field. [**Method/Process**] Building upon the analytical tools of CNKI and EndNote, we applied bibliometric methods to conduct a macro-level analysis of annual trends and core institutions for the topic, followed by thematic content analysis from three perspectives: extraction techniques, datasets and evaluation, and applications. [**Result/Conclusion**] In recent years, term extraction research has made significant progress and constitutes fundamental work in knowledge systems, natural language processing, and information analysis. With the rapid development of natural language processing, extraction technologies have begun to evolve toward deep learning, though the basic theoretical framework of term extraction still requires improvement, particularly regarding evaluation metrics, corpus selection, and effectiveness evaluation methods.

Keywords: term extraction; machine learning; knowledge organization; bibliometrics

Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2020.14.010

With the development of the semantic web, the scope of knowledge content transformation has gradually expanded and accelerated. Multi-channel, multi-format, and linked data heterogeneity in knowledge carriers has become the norm, while user groups increasingly desire effective access to knowledge content. The semantic web is founded on knowledge organization, attempting to achieve semantic interoperability between knowledge units. Taxonomies, ontologies, and knowledge graphs play crucial roles in the semantic web, revealing the connotative semantics between knowledge units, mining extensional associations, enabling data knowledgefication, knowledge ordering, and knowledge servitization, ultimately allowing knowledge to be effectively utilized, disseminated, shared, and enhanced. Terms are linguistic designations of concepts in specific professional fields. Knowledge-intensive systems require large quantities of accurate, standardized terms to describe domain knowledge. In the era of big data, where domain texts, vocabularies, and concepts continue to grow, manually constructing, maintaining, revising, indexing, and describing domain core terms has become a labor-intensive task. Consequently, automatic term extraction (ATE) has emerged as a primary task and foundational work for automatic domain term acquisition research.

Automatic term extraction remains an unresolved problem. Over the years, scholars have developed new methods to meet the growing demands of industry, government archives, and digital libraries for automatic classification and index-

ing of professional documents. These methods typically combine linguistic rules and statistical information: first using language processors to extract candidate terms (e.g., nouns, noun phrases, or n-grams), then applying statistical methods to score candidates based on locally and globally collected features, and finally ranking scored candidates for subsequent selection and filtering. While existing methods have achieved good extraction results, they suffer from two limitations: (1) As is well known, developing a universal "one-size-fits-all" method for any domain is unrealistic. Research shows that the best-performing ATE methods vary by domain and dataset, with potentially significant differences in accuracy across different approaches. (2) Current state-of-the-art techniques typically utilize statistical features such as term frequency to score candidates, neglecting the role of semantic relatedness.

In recent years, machine learning has developed rapidly in the term extraction domain, with its academic value and application prospects continuously explored and exploited, yielding many excellent achievements from theory and models to algorithms and practical applications. By automatically learning optimal combinations of features and cut-off points, machine learning can effectively combine features with broad applicability. Faced with growing and increasingly complex domain data, deep learning—a branch of machine learning—may be a more suitable choice. Term extraction is a complex and challenging task, and the integration of machine learning algorithms has further improved extraction effectiveness and quality, though there remains room for improvement in generalizability and performance. Based on the connection between term extraction and machine learning, this paper conducts a statistical and quantitative analysis of research progress and application status in this field, providing a reference for relevant researchers.

2 Quantitative Analysis of Relevant Literature

Automatic term extraction research has a history of over 20 years. As early as the 1990s, foreign researchers developed a batch of operational term extraction systems, such as the FASTER system and the Terms system, serving information organization and retrieval, text processing, and domain knowledge discovery, organization, and application. Chinese term extraction research started later, primarily improving existing methods based on foreign research while incorporating Chinese language characteristics. Currently, many term service platforms and tools exist, such as the Chinese Thesaurus Service System of the Institute of Scientific and Technical Information of China, the Terminology Knowledge Service Platform Termonline of the China National Committee for Terms in Sciences and Technologies, CNKI' s knowledge element retrieval, OCLC Terminology Services, and SketchEngine.

As literature information resources continue to grow, to comprehensively obtain relevant literature on term extraction research, this paper primarily used CNKI and VIP as Chinese retrieval platforms, selecting "term extraction, term recognition, term acquisition" as primary subject terms for thematic retrieval, then

performing a secondary search using "machine learning, deep learning, neural network, supervised learning, semi-supervised learning, unsupervised learning, conditional random field, support vector machine, maximum entropy, hidden Markov model" as search terms, yielding 290 and 126 documents respectively. After screening by title, abstract, and keywords and removing duplicates, 96 relevant Chinese documents were obtained. For foreign literature, we used the advanced search function of the Web of Science Core Collection database with the following search formula: TS=("term extraction" OR "term recognition" OR "terminology extraction" OR "terminology recognition" OR "term identification" OR "terminology identification") AND ALL=("machine learning" OR "deep learning" OR "neural network" OR "conditional random fields" OR "Support Vector Machine" OR "supervised learning" OR "unsupervised learning" OR "Maximum Entropy" OR "Hidden Markov Model"). Among the 79 retrieved documents, 73 relevant papers were selected after screening.

Considering data availability and research quality, this paper combined the collected Chinese and foreign literature and adopted bibliometric and content analysis methods to analyze the research questions. First, we used CNKI' s data analysis function and EndNote' s Subject Bibliography analysis function to obtain relevant statistical data. Then, we used Excel to analyze annual trends and core research institutions across the total of 169 documents to achieve a macro-level understanding of relevant research. Subsequently, we delved into the literature content and, combined with statistical data, conducted thematic content analysis of machine learning-based term extraction research from three aspects: extraction techniques, datasets and evaluation, and applications.

2.1 Annual Trend Analysis

Automatic term extraction research began in the 1990s. P. Marshall et al. [5] published the earliest conference paper, "Working towards connectionist modeling of term formation," which continued research on connectionist approaches to term recognition and proposed a method using competitive network techniques (winner-take-all algorithms) for automatic term recognition. In Chinese research, Chen Wenliang et al. [7] applied the Bootstrapping machine learning algorithm in 2003 to automatically extract domain vocabulary from large-scale unannotated real corpora. Machine learning-based term extraction research entered a growth phase starting in 2007, followed by a stable period between 2009 and 2013 with an average annual publication rate of about 10 papers. During this period, conditional random fields, support vector machines, domain ontologies, and patent analysis became keywords in term extraction research. Around 2012, deep learning and big data entered rapid development phases, driving advancements in information extraction research such as named entity recognition, keyword extraction, and relation extraction. As one direction of information extraction research, term extraction was also influenced and entered another growth phase starting in 2014. Literature published in the last three years primarily explores the application of neural networks or deep learning

in term extraction research, which to some extent confirms the upward trend shown in Figure 1 [Figure 1: see original paper].

2.2 Core Research Institution Analysis

Research institutions are specialized organizations conducting disciplinary research. Analyzing institutions can reveal the institutional distribution of research directions and help researchers identify sources for academic tracking. By statistically analyzing and deduplicating institutions across the 169 Chinese and English documents, Figure 2 [Figure 2: see original paper] lists 14 major institutions (not excluding co-authorship) with more than 3 publications. Shenyang Aerospace University (formerly Shenyang Institute of Aeronautical Engineering) ranks first with 12 publications, followed by Nanjing University with 10. The highest-publishing foreign institution is the University of Manchester, also the only foreign institution with more than 3 publications. Investigation of foreign term extraction research found that it focuses more on applications, embedded in research on ontologies, knowledge graphs, knowledge systems, and natural language processing. In contrast, Chinese institution research themes are relatively singular and concentrated, focusing more on improving existing technical methods' performance in Chinese, particularly in medical and patent domains, with less application research across different domains.

The 14 institutions collectively published 70 papers, accounting for 42% of the total, showing clear advantages over other research units, especially Shenyang Aerospace University and Nanjing University, which should be key focus institutions for researchers to track. However, these 14 institutions represent a very small proportion of the overall landscape, indicating that the term extraction field still lacks high-productivity, outstanding research institutions.

3 Thematic Analysis

Thematic analysis can reflect the research level and overall status of a field, revealing its current state, hotspots, and development trends. Building on the macro-level analysis results above, we conducted thematic analysis of term extraction-related papers from three main aspects: extraction techniques, datasets and evaluation, and applications.

3.1 Analysis of Extraction Techniques

Traditional term extraction methods include linguistics-based, statistics-based, and multi-strategy hybrid approaches. Linguistics-based methods [8-9] often rely on manual shallow syntactic analysis or domain dictionary construction rules for term extraction, depending on specific languages, domain dictionaries, annotated data, and other prerequisite resources. These suffer from difficulties in rule maintenance and updates, limited application scope, poor scalability and portability, and particularly poor recognition of unregistered terms, resulting

in low precision and recall rates that prevent large-scale application. Statistics-based methods [10-11] utilize distribution statistical properties of terms in domain text corpora, identifying character sequences meeting thresholds or conditions as domain terms using common metrics such as TF-IDF, information entropy, mutual information, and log-likelihood. However, these methods involve large computational costs, easily miss low-frequency terms, and neglect or lack contextual semantic analysis. Different methods can be integrated to combine multiple strategies for improved extraction performance. Based on statistical and linguistic methods, K. T. Frantzi et al. [12] proposed the C-value/NC-value method, which pioneered hybrid strategy research by first using rule templates to obtain candidate term sets, then applying statistical features for filtering. Additionally, Zhou Lang et al. [13] combined substring merging, collocation testing, and domain relevance calculation techniques to improve Chinese phrase-type term extraction system performance.

Traditional term extraction methods can achieve good performance on specific corpora but become increasingly cumbersome against multi-source heterogeneous data and domain intersections. To overcome these limitations, most subsequent research began migrating named entity recognition methods to term extraction research, primarily adopting semi-supervised and supervised hybrid machine learning algorithms and their variants. These focus on semi-automatically or automatically obtaining domain-dependent attributes, specialized text features, and contextual semantic information from domain texts to address the aforementioned problems [14].

To analyze machine learning techniques in automatic term extraction, we statistically analyzed keywords from relevant literature. Keywords are highly refined representations of research content. We primarily performed synonym merging (e.g., “CRF” and “conditional random field”), removed high-frequency keywords unhelpful for thematic research (e.g., “domain term,” “word segmentation,” “research method”), and eliminated overly broad subject terms (e.g., “term extraction,” “machine learning”). Table 1 shows Chinese and English keywords appearing four or more times across the 169 documents. From both Chinese and English keywords, “conditional random field,” “support vector machine,” “neural network,” and “deep learning” appear more than four times, with conditional random fields appearing 53 times total and “deep learning”appearing 23 times, representing frequently applied term extraction techniques in recent years. Based on keyword distribution and considering extraction technique characteristics and development timelines, we categorized extraction methods into statistical machine learning methods and deep neural network methods.

3.1.1 Statistical Machine Learning Methods With the rapid development of machine learning in natural language processing, term extraction research has gradually shifted toward this active area. Statistical machine learning-based term extraction research can be summarized in three directions: model selection, method improvement, and multi-strategy fusion.

(1) Model Selection. Machine learning-based term extraction methods are fundamentally classification approaches that follow two strategies: first identifying term boundaries then classifying, or transforming the problem into sequence labeling.

Classification models are typical statistical learning models in supervised learning, primarily learning classification model weights and parameters from annotated training data to predict new sample categories. P. Lopez et al. [15] compared multiple classification methods for extracting terms from scientific documents: decision trees, support vector machines, and multilayer perceptrons. Zhao Xin [16] utilized large quantities of existing terms to train a term classifier using a maximum entropy model. M. Shirakawa et al. [17] proposed an extended naive Bayes model to extract key terms from texts for classifying noisy short texts. W. Zeng et al. [18] used SVM for term extraction from new energy vehicle domain patents and literature, with experimental results confirming machine learning' s effectiveness in term extraction. Table 2 summarizes the functions and characteristics of classification algorithms used for term extraction, which have also achieved great success in natural language processing fields such as text classification, speech recognition, and image understanding.

Sequence labeling models can solve common natural language problems including part-of-speech tagging, named entity recognition, and semantic role labeling. Unlike general classification models, sequence labeling models treat text as a sequence, using labeling methods such as BIO, BIEO, and BMEO for term recognition, representing a highly effective current approach. Table 3 introduces two commonly used sequence labeling models in term extraction: Hidden Markov Model (HMM) and Conditional Random Field (CRF). H. S. Pan et al. [19] proposed using HMM to extract new terms from academic literature for Chinese lexicon construction. Cen Yonghua et al. [20] applied HMM to computer domain corpora, achieving an F-value of 89.75%. Compared with HMM, CRF offers greater advantages, avoiding label bias problems. Zhang Chengzhi [21] proposed an integrated term extraction strategy based on CRF. D. Zheng et al. [22] used discrete term features as CRF template attributes, adjusting feature templates from multiple perspectives including word itself, word position in compound terms, text semantic information, information entropy, and TF/IDF, achieving good results in domain term recognition.

(2) Method Improvement. Machine learning has achieved rapid success in term extraction research. To design better-performing term extraction methods, researchers have improved existing models to enhance recognition effectiveness and computational efficiency, such as Q. Zhan et al.' s [23] cascaded conditional random field model. Improvement research is more common in Chinese term extraction because classical models mostly target English and cannot be directly applied to Chinese. Adjusting and optimizing classical models can more effectively recognize terms in Chinese texts.

A problem with statistical machine learning methods is their reliance on domain-specific feature engineering. To improve algorithm accuracy, expert knowledge

(experience) and "luck" are required—that is, the process of manual feature selection is random and uncontrollable, making large-scale popularization difficult. Therefore, another approach to improving term extraction effectiveness is selecting better feature representations [24]. Common features in term extraction tasks include morphological, lexical, and syntactic information. Morphological features include word forms and affixes, while lexical and syntactic features include word length, part-of-speech, shallow syntactic parsing, and dependency parsing. Considering Chinese particularities, character-level features such as radicals and strokes are also used to assist term extraction. Meanwhile, various external knowledge sources like dictionaries, Wikipedia, synonym forests, HowNet, and CN-Probase can improve recognition performance.

(3) Mixed Strategy. Mixed strategy methods can effectively reduce computational complexity while fully utilizing contextual semantic information for domain text analysis, improving recognition performance to some extent. C. Y. Chi et al. [25] combined one-hot encoded Brown clustering with HMM for unsupervised learning on unlabeled corpora. Additionally, Huang Han et al. [26] proposed a CRF model combined with active learning, iteratively improving classifier efficiency to achieve precision and recall rates above 90%.

3.1.2 Deep Neural Network Methods Since 2012, the surge in deep neural networks and deep learning development has achieved fruitful results in speech recognition, image recognition, and computer vision. Particularly, word embedding-based semantic representation methods such as Word2Vec, fastText, GloVe, ELMo, BERT, and XLNet have solved data sparsity problems caused by high-dimensional vector spaces. These methods can use word embeddings to obtain feature representations rich in semantic information from heterogeneous texts, injecting strong development momentum into domain-specific term extraction. The advantage of deep learning is its ability to use various deep neural network models or algorithms to automatically learn features from domain texts, avoiding heavy and time-consuming feature engineering. The feature learning process is independent of manual, domain, and language factors, thus offering strong portability, reusability, and scalability [14].

To address over-reliance on feature engineering and poor generalization performance on complex problems in existing machine learning methods, recent research has begun exploring term extraction based on deep neural network methods. R. Chalapathy et al. [27] found that traditional machine learning methods heavily depend on manual features and domain-specific resources, proposing a BLSTM-CRF model to extract medical concepts from clinical data, achieving better results than HMM, CRF, and other ATE algorithms. R. Wang et al. [28] introduced a weakly supervised bootstrapping method using two deep learning classifiers for term extraction, effectively alleviating problems of manual feature selection and lack of labeled data.

With continuous deep learning development, researchers have proposed optimization mechanisms. The attention mechanism essentially simulates human

brain focus characteristics that concentrate attention on specific key items while ignoring non-critical items at particular moments. Ma Jianhong et al. [29] proposed a BLSTM-CRF domain term extraction model based on the Attention mechanism, achieving 86% precision. Transfer learning migrates labeled data or knowledge structures from related domains to complete or improve target domain/task learning effectiveness. Liu Yufei et al. [30] introduced deep transfer learning ideas, using the BiLSTM model for cross-domain migration to effectively identify technical terms, solving the problem of limited patent literature annotations. Domain knowledge is crucial for term extraction in domain-specific corpora but difficult to obtain from limited corpora. Using domain facts derived from knowledge bases such as Wikipedia and Baidu Baike to learn term features through distant supervision can achieve broader coverage than existing methods [31].

Current deep learning methods applied in term extraction are transplanted from named entity recognition research based on domain characteristics, thus facing the same problems of lacking large-scale standardized texts, annotated corpora, and basic lexicons. Research results show significant improvement in extraction accuracy, but the ideal peak has not yet been reached. On the deep learning technology foundation, improving extraction efficiency and more effectively utilizing limited annotated data are worthwhile research directions in term extraction, such as fine-tuning pre-trained models (BERT, XLNet) on domain corpora.

3.2 Dataset and Evaluation Analysis

3.2.1 Dataset Analysis Automatic term extraction is a fruitful research area but still faces significant obstacles in datasets and evaluation, requiring manual term annotation—a difficult and arduous task. The lack of clear distinction between terms and general language leads to low inter-annotator agreement, increasing annotation ambiguity. With the continuous development of machine learning and deep learning methods, the demand for annotated datasets has become increasingly urgent, not only for evaluation but also because "one of the main problems in applying machine learning or deep learning to ATE is the availability of reliable training data."

Through reading and summarizing experimental sections of papers, datasets are mainly divided into public datasets and research-specific datasets. Public datasets are publicly available annotated datasets with broad applicability, including GENIA, ACLRD-TEC, FAO, etc. Table 4 shows statistical information for commonly used datasets. Among them, GENIA is the most frequently used dataset for ATE evaluation, serving as a semantic annotation dataset for biomedical text mining. The ACL dataset is specifically designed for ATE evaluation in the NLP field, based on the assumption that having a dataset allowing NLP researchers to become domain experts themselves would be a tremendous advantage. In addition to the datasets in the table, there are smaller public resources such as TTCm and TTCw [2]. The TTCw corpus contains 103 full-text articles on wind energy, while TTCm contains 37 full-text articles on mobile

technology.

We found that public datasets are primarily English-based, while Chinese research mainly constructs domain datasets manually for research purposes. Huang Han et al. [26] used judicial documents as research objects, crawling 61,515 judicial documents from "China Judgments Online," and after data cleaning, manually annotated five categories of terms including charges, penalties, legal principles, legal concepts, and legal provisions. To extract new energy vehicle domain terms, Ma Jianhong et al. [29] manually annotated 1,126 patent texts and validated them in the CAI innovation tool. Multilingual term extraction research is an emerging field; R. A. Terryn et al. [32] collected corpora in three languages (English, French, and Dutch) and four domains (corruption, dressage, heart failure, and wind energy), designing annotation schemes. Research-specific datasets cover broad domains including finance, military, library and information science, scientific and technical literature, patents, and web texts.

3.2.2 Evaluation Analysis Traditional ATE evaluation methods compare results with manual annotations, calculating precision (number of actual candidate terms), recall (number of correctly extracted terms), and F-value (harmonic mean of precision and recall). For example, Huang Han et al. [26] evaluated legal term recognition effectiveness using precision P, recall R, and F-value. These three metrics cannot comprehensively reflect extraction quality, being closely related to noise (incorrectly extracted terms) and silence (unextracted terms). In addition to these metrics, receiver operating characteristic curve (ROC) is also an evaluation method but is less common in term extraction. Since these metrics only measure performance, some researchers argue that more comprehensive evaluation protocols are necessary. As early as 1996, M. C. L' Homme et al. [33] broadly defined five pre-evaluation criteria to supplement the above metrics. In other work, V. A. Sauron [34] proposed a quality model that calculates not only precision or recall but also measures applicability, reliability, usability, maintainability, and portability. Similarly using only P, R, and F, Zhao Hong et al. [35] explored the impact of training corpus scale on extraction results, calculating extraction performance at 20%, 40%, 60%, and 80% training set proportions. D. Inkpen et al. [36] considered mixing multiple evaluation strategies and designed tools to facilitate comparative evaluation of ATE systems.

3.3 Application Analysis

As shown in Table 5 , machine learning-based term extraction applications include knowledge organization, natural language processing, information analysis, and others. In library and information science, applications mainly involve thesaurus and ontology construction, scientific and technical information analysis, and patent term extraction to support information system construction and services. Term extraction is a fundamental task for data and knowledge acqui-

sition and a preprocessing step for many complex natural language processing tasks such as information retrieval, machine translation, text mining, relation extraction, etc. Other applications refer to domain-specific term extraction tasks based on research objectives, covering domains including finance, military, law, medicine, commerce, and agriculture. Summarizing application situations helps researchers understand the field' s current status, identify research directions, and uncover research value.

3.3.1 Thesaurus Maintenance and Update In biomedicine, computer science, natural sciences, and other fields, new terms emerge with new technologies and knowledge. To promote sharing and utilization of domain thesaurus resources, thesaurus maintenance and updates are imperative. M. Ikeda et al. [37] used machine learning for candidate term extraction from the perspective of extending multiple thesauri, then added corresponding unregistered terms to thesauri based on grammatical information. The use of thesauri and taxonomies to obtain scientific and technical information in scientometrics has long been of interest. T. Kawamura et al. [38] proposed using Word2Vec tools to obtain domain-related new concepts and terms from abstracts in advanced technology fields to extend domain thesauri and keep abreast of latest trends in various scientific and technological activities. Song Peiyan et al. [39] studied thesaurus construction methods under the semantic web environment, proposing the use of machine learning methods to automatically extract terms from corpora and literature resources to construct initial term sets. Additionally, for thesauri such as MeSH and GeoRef Earth Science Thesaurus, machine learning-based term extraction will play a crucial role in providing comprehensive information retrieval services under the big data background.

3.3.2 Domain Ontology Construction Domain ontologies are explicit formal specifications of shared conceptual models, using recognized term sets and term relationships to reflect knowledge and knowledge structures within a domain, playing important roles in semantic information interaction and standardized information description. Terms are basic elements of domain ontology construction, and term extraction is the most fundamental and crucial step in ontology learning. To improve ontology construction efficiency and reduce costs, B. Omelayenko [40] early used machine learning methods for term extraction, ontology merging, updating, and instance acquisition. Li Lishuang [41] proposed a domain term extraction method combining conditional random fields and active learning, achieving a certain degree of automation in the ontology construction process and providing a good method for manufacturing enterprise knowledge management modeling. To construct domain academic ontologies, Jiang Ting [42] used a cascaded conditional random field combined with C-value and rules to extract different term types.

3.3.3 Natural Language Processing Machine learning-based term extraction can also be applied in natural language processing. R. Gaizauskas et al. [43]

introduced a multi-component system BiTES for automatically extracting bilingual term pairs from Web sources, first extracting terms from monolingual corpora automatically, then aligning extracted terms from comparable documents or parallel corpora. G. Huang et al. [44] found that parentheses on web pages contain substantial term translation knowledge. To improve extraction recall, they proposed a maximum entropy-based term recognition system TermExt and used supervised machine learning methods for machine translation of extracted terms, achieving an 11% recall improvement over the baseline. In information retrieval, N. T. W. Khin et al. [45] proposed a Web query classification algorithm-based IR system including domain term extraction, Web query classification, and relevant query retrieval. The Unified Medical Language System (UMLS) integrates over 150 medical thesauri and is widely used for retrieving and mining Internet literature. IEEE' s top-level ontology SUMO also attempts to integrate knowledge organization tools including thesauri to provide more comprehensive knowledge retrieval services.

3.3.4 Information Analysis In the big data environment, information analysis for scientific and technological information monitoring and knowledge acquisition has become increasingly important. Scientific and technical terms can represent scientific and technological concepts and express core content of scientific and technological data, forming an important component of scientific and technological data information analysis. Zeng Wen et al. [46] introduced deep learning-based scientific and technical term extraction methods and conducted experimental analysis and conclusions on scientific and technical datasets. Zeng Wen, Che Yao, et al. [47] proposed methods for scientific and technical big data information analysis services from the perspective of scientific and technical big data, designing and developing a Chinese term extraction method integrating multiple extraction algorithms. Experiments showed this method can assist information researchers in data processing and analysis to some extent. Theoretical terms are fundamental for large-scale literature content analysis and deep cross-disciplinary knowledge transfer revelation. Zhao Hong et al. [35] constructed a deep learning model for theoretical term extraction. Patent literature analysis can determine domain technology hotspots, predict technology development trends, and help researchers gain inspiration and reference, with patent literature terms providing structured knowledge as a key component of patent literature analysis [48-49].

Conclusion

This paper employs bibliometric analysis and content analysis to examine papers related to machine learning techniques in the Web of Science, CNKI, and VIP databases under relevant subject terms. Through bibliometric analysis, we conducted macro-level analysis of external characteristics of the dataset, including annual trends and core institutions. We found that term extraction research remains in a growth phase along with related fields' rapid development, with academic tracking possible by following core institutions such as "Shenyang

Aerospace University" and "Nanjing University." Subsequently, we conducted thematic analysis of 169 Chinese and English documents from three aspects: extraction techniques, datasets and evaluation, and applications, reaching the following conclusions:

(1) The introduction of statistical machine learning methods has significantly advanced term extraction technology, but model recognition performance heavily depends on annotated corpus quality and feature engineering. In recent years, the deep learning boom has further promoted term extraction research development, as deep learning can automatically learn features and reduce dependence on domain knowledge. However, current research results show term extraction remains far from solved and continues to be a challenging research field. In the big data environment, machine learning and deep learning will be the most effective term extraction methods. To further improve model accuracy, many aspects merit consideration, such as adopting hybrid methods, combining domain knowledge bases, and using pre-trained models.

(2) Datasets and evaluation methods for automatic term extraction are crucial for quantifying state-of-the-art performance and should include text corpora, gold standards, and evaluation metrics. A. R. Terryn et al. [32] provided some standard datasets and annotation strategies across several domains and languages. Section 3.3.2 also introduced domestic and international datasets and evaluation methods, which are invaluable for correctly evaluating term extraction models. In multi-source heterogeneous data environments, datasets and evaluation still face significant obstacles, and the term extraction theoretical framework needs improvement, including corpus selection, evaluation metrics, and effectiveness evaluation methods.

(3) Machine learning-based term extraction technology is fundamental and important work for knowledge systems, natural language processing, information analysis, and other research fields, with high practical value. Applications are not limited to those mentioned in Section 3.3; more applications across different domains await further exploration by researchers. Indeed, with data becoming massive, heterogeneous, and complex, machine learning and deep learning will play increasingly important roles in term extraction.

This study has limitations, including inability to guarantee comprehensive and accurate literature collection, errors in bibliometric analysis, and insufficient depth in thematic and application domain analysis. We hope this research accurately reflects the current status of machine learning-based term extraction research and welcome criticism and corrections from experts and scholars.

References

[1] Terminology work—Principles and methods [J]. Terminology Standardization

and Information Technology, 2003(1): 45-48.

[2] ZHANG Z, GAO J, CIRAVEGNA F. Semre-rank: improving automatic term extraction by incorporating semantic relatedness with personalised pagerank [J]. ACM transactions on knowledge discovery from data, 2018, 12(5): 1-41.

[3] ASTRAKHANTSEV N. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in scala [J]. Language resources and evaluation, 2018, 52(3): 853-872.

[4] CASTELLVÍ M T C, BAGOT R E, PALATRESI J V. Automatic term detection: a review of current systems [C]// Proceedings of the international conference on computational terminology. Berlin: Springer, 2001: 53-88.

[5] MARSHALL P, BANDAR Z. Working towards connectionist modeling of term formation [C]// Proceedings of the international conference on computational intelligence. Heidelberg: Springer, 1999: 522-529.

[6] BENGIO Y. A connectionist approach to speech recognition [J]. International journal of pattern recognition and artificial intelligence, 1993, 7(4): 647-667.

[7] CHEN Wenliang, ZHU Jingbo, YAO Tianshun, et al. Bootstrapping-based domain vocabulary automatic acquisition [C]// Proceedings of the 7th National Joint Conference on Computational Linguistics. Beijing: Tsinghua University Press, 2003: 67-72.

[8] KAUSHIK N, CHATTERJEE N. A practical approach for term and relationship extraction for automatic ontology creation from agricultural text [C]// Proceedings of the 2016 international conference on information technology. Bhubaneswar: IEEE, 2016: 241-247.

[9] STANKOVIĆ R, KRSTEV C, OBRADOVIĆ I, et al. Rule-based automatic multi-word term extraction and lemmatization [C]// Proceedings of the 10th international conference on language resources and evaluation. Portorož, Slovenia: European Language Resources Association, 2016: 507-514.

[10] DU L, LI X, LIN D. Chinese term extraction from webpages based on expected pointwise mutual information [C]// Proceedings of the 2016 12th international conference on natural computation, fuzzy systems and knowledge discovery. Changsha: IEEE, 2016: 1647-1651.

[11] LI Lishuang, WANG Yiwen, HUANG Degen. Term extraction research based on information entropy and word frequency distribution changes [J]. Journal of Chinese Information Processing, 2015, 29(1): 82-87.

[12] FRANTZI K T, ANANIADOU S, TSUJII J. The c-value/nc-value method of automatic recognition for multi-word terms [C]// Proceedings of the international conference on theory and practice of digital libraries. Berlin: Springer, 1998: 585-604.

[13] ZHOU Lang, SHI Shumin, FENG Chong, et al. Chinese term extraction method based on multi-strategy fusion [J]. Journal of the China Society for

Scientific and Technical Information, 2010(3): 460-467.

[14] WANG Sili, ZHU Zhongming, LIU Wei, et al. Research on automatic domain ontology concept acquisition methods based on deep learning [J]. Information Studies: Theory & Application, 2019(10): 1-13.

[15] LOPEZ P, ROMARY L. HUMB: automatic key term extraction from scientific articles in GROBID [C]// Proceedings of the 5th international workshop on semantic evaluation. Los Angeles: Association for Computational Linguistics, 2010: 248-251.

[16] ZHAO Xin. Design and implementation of Chinese term extraction system based on maximum entropy [D]. Xi' an: Xidian University, 2012.

[17] SHIRAKAWA M, NAKAYAMA K, HARA T, et al. Wikipedia-based semantic similarity measurements for noisy short texts using extended naive Bayes [J]. IEEE transactions on emerging topics in computing, 2015, 3(2): 205-219.

[18] ZENG W, LI X, LI H. Study on Chinese term extraction method based on machine learning [C]// Proceedings of the international conference of pioneering computer scientists, engineers and educators. Singapore: Springer, 2018: 128-135.

[19] PAN H S, ZHAO J Y. Combining syntactic information with HMM for term extraction [C]// Proceedings of the 2015 2nd international conference on information science and control engineering. Washington, DC: IEEE Computer Society, 2015: 170-173.

[20] CEN Yonghua, HAN Zhe, JI Peipei. Chinese term recognition based on hidden Markov model [J]. Data Analysis and Knowledge Discovery, 2008, 24(12): 54-58.

[21] ZHANG Chengzhi. Integrated term extraction research based on multi-layer termhood [J]. Journal of Intelligence, 2011, 30(3): 275-285.

[22] ZHENG D, ZHAO T, YANG J. Research on domain term extraction based on conditional random fields [C]// International conference on computer processing of oriental languages. Heidelberg: Springer, 2009: 290-296.

[23] ZHAN Q, WANG C. A hybrid strategy for Chinese domain-specific term extraction [C]// Proceedings of the 2015 11th international conference on semantics, knowledge and grids. Washington, DC: IEEE Computer Society, 2015: 217-221.

[24] RIGOUTSTERRYN A, DROUIN P, HOSTE V, et al. Analysing the impact of supervised machine learning on automatic term extraction: HAMLET vs TermoStat [C]// Proceedings of the international conference on recent advances in natural language processing. Varna, Bulgaria: INCOMA Ltd., 2019: 1012-1021.

[25] CHI C Y, ZHANG Y. Information extraction from Chinese papers based on hidden Markov model [J]. Advanced materials research, 2013, 846: 1291-1294.

[26] HUANG Han, WANG Hongyu, WANG Xiaoguang. Conditional random field model combined with active learning for automatic legal term recognition [J]. Data Analysis and Knowledge Discovery, 2019, 3(6): 66-75.

[27] CHALAPATHY R, BORZESHIEZ, PICCARDI M. Bidirectional LSTM-CRF for clinical concept extraction [C]// Proceedings of the clinical natural language processing workshop. Osaka: The COLING 2016 Organizing Committee, 2016: 7-12.

[28] WANG R, LIU W, MCDONALD C. Featureless domain-specific term extraction with minimal labelled data [C]// Proceedings of the Australasian Language Technology Association workshop 2016. Melbourne: Australasian Language Technology Association, 2016: 103-112.

[29] MA Jianhong, ZHANG Yamei, YAO Shuang, et al. New energy vehicle domain term extraction based on BLSTM_{{{Attention}}_{{CRF}}} model [J]. Application Research of Computers, 2019(5): 1-34.

[30] LIU Yufei, YIN Li, ZHANG Kai, et al. Technical term identification based on deep transfer learning: a case study of CNC systems [J]. Journal of Intelligence, 2019, 38(10): 168-175.

[31] ALFARO N E, DAVIS J. Unsupervised learning of an is-a taxonomy from limited domain-specific corpus [C]// Proceedings of the 24th international joint conference on artificial intelligence. Buenos Aires: AAAI Press, 2015: 1434-1441.

[32] TERRYN A R, HOSTE V, LEFEVERE. In uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora [J]. Language resources and evaluation, 2020, 54(1): 1-24.

[33] L' HOMME M-C, BENALI L, BERTRAND C, et al. Definition of an evaluation grid for term-extraction software [J]. Terminology international journal of theoretical and applied issues in specialized communication, 1996, 3(2): 291-312.

[34] SAURON V A. Tearing out the terms: evaluating term extractors [C]// Proceedings of translating and the computer 2002. London: Aslib, 2002: 1-18.

[35] ZHAO Hong, WANG Fang. Deep learning model and self-training algorithm for theoretical term extraction [J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(9): 923-938.

[36] INKPEN D, PARIBAKHT T S, FAEZ F, et al. Term evaluator: a tool for terminology annotation and evaluation [J]. International journal of computational linguistics and applications, 2016, 7(2): 145-165.

[37] IKEDA M, YAMAMOTO A. Extending various thesauri by finding synonym sets from a formal concept lattice [J]. Information and media technologies, 2017(12): 240-266.

[38] KAWAMURA T, KOZAKI K, KUSHIDA T, et al. Expanding science and technology thesauri from bibliographic datasets using word embedding [C]//

Proceedings of the 2016 IEEE 28th international conference on tools with artificial intelligence. San Jose: IEEE, 2016: 857-864.

[39] SONG Peiyan, CHEN Baixue, WANG Xing. Research on thesaurus construction methods under the semantic web environment [J]. Information Science, 2018, 36(2): 14-17.

[40] OMELAYENKO B. Learning of ontologies for the Web: the analysis of existent approaches [C]// Proceedings of the international workshop on Web dynamics. London: WebDyn@ICDT, 2001: 16-21.

[41] LI Lishuang. Research on term and relationship extraction methods in domain ontology learning [D]. Dalian: Dalian University of Technology, 2013.

[42] JIANG Ting. Research on domain ontology learning and academic resource semantic annotation [D]. Nanjing: Nanjing University, 2017.

[43] GAIZAUSKAS R, PARAMITA M L, BARKER E, et al. Extracting bilingual terms from the Web [J]. Terminology international journal of theoretical and applied issues in specialized communication, 2015, 21(2): 205-236.

[44] HUANG G, ZHANG J, ZHOU Y, et al. Learning from parenthetical sentences for term translation in machine translation [C]// Proceedings of the 9th SIGHAN workshop on Chinese language processing. Taipei: Association for Computational Linguistics, 2017: 37-45.

[45] KHIN N T W, YEEN N N. Query classification based information retrieval system [C]// Proceedings of the 2018 international conference on intelligent informatics and biomedical sciences. Bangkok: IEEE, 2018: 151-156.

[46] ZENG Wen, LI Hui, XU Hongjiao, et al. Application research of deep learning technology in scientific literature data analysis [J]. Information Studies: Theory & Application, 2018, 41(5): 110-113.

[47] ZENG Wen, CHE Yao, ZHANG Yunliang, et al. Research on methods and tools serving scientific big data information analysis [J]. Information Science, 2019, 37(4): 92-96.

[48] YU Yan, ZHAO Naxun. Patent topic discovery method incorporating term knowledge [J]. Library and Information Service, 2018, 62(21): 118-126.

[49] WANG Jian, YIN Xu, LYU Xueqiang, et al. Patent literature domain term extraction method based on CRFs [J]. Computer Engineering and Design, 2019, 40(1): 279-284.

Author Contributions:
Qiu Keda: Literature research, paper writing;
Ma Jianling: Topic suggestion, paper revision and polishing.

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