Research on Keyword-Based Element Extraction for Chinese Patent Retrieval

1st Zeying Jin
School of Information Science
Beijing Language and Culture University
Beijing, China
zeyi_jin@126.com

2nd Zhaoyong Yang
School of Information Science
Beijing Language and Culture University
Beijing, China
yangzhaoyong_blcu@163.com

3rd Gongbo Tang
School of Information Science
Beijing Language and Culture University
Beijing, China
gongbo.tang@blcu.edu.cn

4th Tingchao Liu
School of Information Science
Beijing Language and Culture University
Beijing, China
liutingchao@hotmail.com

5th Endong Xun*
Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture University
Beijing, China
edxun@126.com
*Corresponding Author

***Abstract*—Patent retrieval is a critical step in patent analysis. Retrievable elements play a key role in constructing search queries and performing accurate searches, and most retrievable elements are created manually. However, the increment of patent applications each year has brought a huge burden on manual extraction of retrievable elements and patent examination, raising the urgent need of automated solutions. As keywords serve as an effective way of expressing retrievable elements in patent retrieval, we explore the automatic extraction of keyword-based retrievable elements from Chinese patent application texts in this study. We employ various keyword extraction methods, including large language model based methods, to identify retrievable elements within these texts. Our experimental results have shown that these methods can effectively extract keywords as retrievable elements from Chinese patent applications, which benefits to manual patent searching and patent examinations.**

***Index Terms*—patent retrieval, retrievable elements, keyword extraction**

I. Introduction

Patents are fundamental components of intellectual property, protecting innovations by granting exclusive rights to their owners. With the increasing awareness of intellectual property protection, patent applications have a dramatic increase, which makes patent documents being the crucial sources of technical documents. These documents contain over 90% of the world's latest technological advancements [10]. Businesses and individuals need to navigate extensive patent documents to understand technological trends and secure competitive advantages, while patent offices must find prior art to assess an invention's patentability and define its scope of protection. Both groups need efficient patent retrieval.

Patent retrieval, a subfield of information retrieval, is crucial in patent analysis and examination [1]. It relies on the content of application texts, which are often multi-modal, multilingual, semi-structured, and rich in metadata. The complexity and length of these texts, together with specialized terminologies, make patent retrieval more challenging compared to conventional text retrieval, particularly in the context of automation. In addition, manual patent examination is time-consuming and resource-intensive, which results in application backlogs and delays. With the development of natural language processing (NLP) models, utilizing NLP techniques to automate aspects of the retrieval process can expedite examinations and reduce complexity.

However, current studies focus on extracting specialized terminology, named entities, and semantic relationships, due to the specific nature of patent texts. These methods may not be suitable for extracting retrievable elements required for patent examination. Thus, we apply varioius keyword extraction methods to identify retrievable elements in patent texts in this study, to investigate the effectiveness of these methods. More specifically, we utilize various unsupervised and supervised techniques to get retrievable elements across multi-domain patent texts, and evaluate the retrieval performance. Our experimental results show that keyword extraction methods can effectively extract retrievable elements from Chinese patent applications, which benefits to manual patent searching and patent examinations.

II. Preliminaries

A. Patent Application Texts

Patent documents detail inventions and encompass texts generated throughout a patent's lifecycle—from application and publication to examination, grant, invalidation, and termination [13]. Among these, patent application texts are of paramount importance as they initiate the patent lifecycle and serve as key references during examination. Examiners rely on the technical content of these texts as the foundation for retrieval, assessing the novelty and inventiveness of the patent by comparing it with existing technologies.

Supported by the NSFC (No. 62076038); achievements of the Project of Intelligent International Chinese Education at Beijing Language and Culture University.

The Patent Law mandates that application texts for inventions or utility models include a request for granting a patent, a specification, an abstract, and claims. The cover page contains technical information such as the title, kind codes, and abstract, as well as legal information such as the inventors, application date, and legal status. Claims and descriptions delineate the invention's scope of protection. According to Article 64, Paragraph 1 of the Patent Law, ”the scope of protection of an invention or utility model patent shall be determined by the content of its claims, with the description and drawings being used to interpret the content of the claims” [1]. Claims are the primary focus during examination and retrieval, as they outline the technical solution for which protection is sought. Given the specialized, semi-structured, and extensive nature of application texts, experimental analyses typically focus on the claims, supplemented by descriptions when necessary.

B. *Retrievable Elements*

Retrievable elements, as referenced in [1], represent the core conceptual aspects of a technical solution and serve as criteria or factors for narrowing down search parameters [19]. This concept is abstract and not confined to a specific part of speech or category, necessitating specific expressions. Typically, there are two primary means of articulation: classification codes and keywords.

During the initial examination phase, patents are classified based on their technical subjects using the International Patent Classification (IPC) system, which assigns one or more classification codes to each patent application. While the use of classification codes for retrieval offers speed advantages, the broad categorization of certain patent classification codes may introduce significant noise into search results. Moreover, in cases of interdisciplinary applications where classification codes may lack precision, the inclusion of keywords for initial retrieval becomes imperative [4]. Although intelligent semantic retrieval technology enables direct searches for relevant text fragments even in instances where keywords are insufficient, it is predominantly utilized for ranking search results and assisting keyword and classification code-based searches.

Thus, keywords, as one of the retrievable elements, play a critical role in ensuring the precision of patent retrieval and the reliability of search outcomes.

C. *Patent Retrieval*

Depending on the retrieval objectives, patent searches can be classified into two types: prior art searches, which evaluate novelty and inventiveness, and technical topic searches, which focus on patent analysis. The former focuses on retrieving existing technologies and is commonly used in the patent examination process for novelty searches. The latter aims to find documents that describe the same technical topics, as in thematic searches. Although the goals differ, leading to variations in specific details of the search process, the overall steps are generally similar, as illustrated in Fig. 1.

In a patent search, the process begins with understanding the technical concept delineated by the independent claims, which serves as the subject of the search. Subsequently, based on the technical solution outlined in the claims, retrievable elements are extracted from aspects such as the technical field, technical problems, technical means, or effects. These retrievable elements are then combined into a search query, which is executed in a selected patent database. After a preliminary search, the search strategy may need to be refined based on the results and intended objectives. This might involve modifying the expression or combination of retrievable elements. The process is repeated with these adjustments until the termination criteria for the search are met.

In Section IV, we will follow this basic workflow in our experiments.

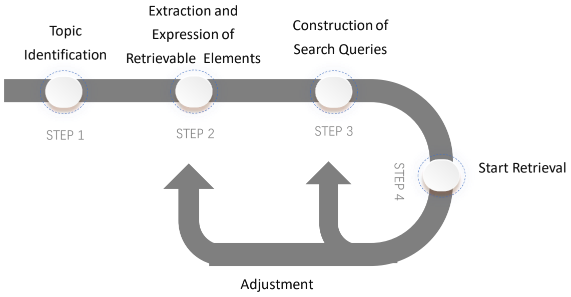
Topic Identification
Extraction and Expression of Retrievable Elements
Construction of Search Queries
STEP 1
STEP 2
STEP 3
STEP 4
Start Retrieval
Adjustment

Fig. 1. Basic process of patent retrieval.

III. Related Work

A. *Existing Research on Patent Information Extraction and Associated Challenges*

The extraction of patent information is fundamental for effective patent retrieval and mining. Due to the lengthy, complex, and highly specialized nature of patent texts, and the variability in terminology and relational expressions across different fields, research in patent information extraction has primarily targeted specific domains, such as shipbuilding, or focused on specific types of data, such as terminology, named entities, and semantic relationships [5]. These specialized studies have paved the way for further tasks such as patent representation, retrieval, mining, and generation.

However, in the context of patent examination and retrieval, the types and expressions of retrievable elements are diverse and not restricted to specific parts of speech or categories. As a result, existing techniques for patent information extraction are not entirely suitable for extracting these retrievable elements.

In this paper, we explore the feasibility of using more general keyword extraction methods for extracting retrievable elements from patent application texts across multiple domains. We conduct experiments to compare various extraction methods and the granularity of their results, evaluating their performance against gold standard and search engine results. This study aims to assess the potential of computer-assisted extraction of retrievable elements, providing insights for future research and practical applications in patent retrieval.

B. *Keyword Extraction Methods*

The essence of keyword-based retrievable elements lies in the task of keyword extraction. In this section, we will summarize typical keyword extraction methods, as shown in Tab. I. Research on keyword extraction can be divided into two categories: unsupervised and supervised methods. Unsupervised methods can be further divided into several types: statistics-based, graph-based, neural network-based, embedding-based models with pre-trained language models (PLM), and prompt-based methods with large language models (LLM).

Supervised methods typically employ an end-to-end approach that integrates both the selection and ranking processes of keywords, optimizing both stages simultaneously. However, supervised methods require annotated corpora, making it difficult to generalize to new domains. Additionally, as language models evolve, the increasing number of parameters leads to more complex and resource-intensive training processes. Thus, the adoption of supervised keyword extraction methods involves balancing factors such as data availability, computational capability, and training duration.

TABLE I
KEYWORD EXTRACTION METHODS

Unsupervised	Statistics-based Methods	TF-IDF [16], RAKE [15],YAKE [3]
	Graph-based Methods	TextRank [17], TopicRank [2], PositionRank [7]
	Embedding-based Methods	Key2vec [12], KeyBert [8], PromptRank [9]
	Prompt-based Methods	ChatGPT, ChatGLM[1], Qwen[2]
Supervised	Pre-trained Language Models	ELMo, Bert, Transformer
	Large Language Models	Efficient-tuning
		Fine-tuning

IV. Experiments on Extracting Chinese Patent Retrievable elements Based on Keywords

To investigate the effectiveness of various keyword extraction methods for extracting retrievable elements from Chinese patents, we selected one or two representative methods from each category mentioned in Section III-B. We conducted experiments using a dataset of Chinese patent application texts. The following sections detail the dataset, evaluation metrics, experimental setup, and results.

A. *Datasets*

The dataset is divided into two parts: a test set and a retrieval dataset. The test set is used to extract retrievable elements, while the retrieval dataset is used to evaluate the quality of these elements. We applied various keyword extraction methods to the test set to extract retrievable elements and evaluated their quality from two perspectives:

Quality of Extraction: This is measured using the F1-score, which assesses the accuracy and completeness of the extracted retrievable elements.

Retrieval Performance: The extracted retrievable elements are used to construct search queries, which are then input into a search engine to retrieve relevant documents from the retrieval dataset. The quality of the retrievable elements is evaluated based on the relevance of the search results.

1) Retrieval Dataset: To validate the effectiveness of keyword extraction methods across multiple domains of Chinese patent application texts, we constructed a database for retrieval experiments. This database comprises over 25,000 Chinese patent application texts covering the eight main sections (A-H) of the International Patent Classification (IPC). Henceforth referred to as the retrieval dataset, it serves as the basis for measuring the quality of retrievable element extraction.

We conducted a statistical analysis of the distribution of Chinese patent application texts across the eight main sections (A-H) of the IPC from 2014 to 2020. The proportions of these texts within each section are summarized in Tab. II.

For experimental convenience, we randomly sampled 25,696 patent application texts from the six-year period according to their natural distribution. This sampling strategy aims to simulate the composition of a real patent database used in patent retrieval processes.

The dataset includes both Chinese invention patents and utility model patents. Each patent application text contains various metadata such as patent type, filing date, application number, kind code, and inventors, as well as sections such as title, abstract, claims, description, and illustration descriptions, detailing the invention content.

TABLE II
DISTRIBUTION OF THE EIGHT CATEGORIES IN THE TEST SET

Categories	A	B	C	D	E	F	G	H	Total
Proportion /%	18.18	23.04	15.33	1.97	4.31	9.84	14.05	13.27	100

2) Test Set: The test set consists of 764 Chinese invention and utility model patents spanning the eight main sections (A-H) of the IPC. These patents have all been examined by patent examiners and rejected due to prior art references that impacted their novelty or inventive step.

Each patent in the test set includes application text (kind codes, title, abstract, claims, description, and illustration descriptions) and retrievable elements (gold standard) provided by the examiner in patent examination reports, along with the document numbers of related prior art references. These related documents, identified using the keywords, exhibit varying degrees of relevance to the rejected patent application. Each patent has 1 to 5 related documents, all of which are present in the retrieval dataset described in Section IV-A1.

This paper investigates the application of keyword extraction methods in the patent domain. Therefore, the retrievable elements in the test set specifically refer to search elements expressed as ”keywords” (hereafter referred to as such). The experiments do not involve the use of ”kind codes” for retrieval. Since keyword extraction generally involves extracting keywords directly from the original text, the test set only includes keywords that appear in the original text and can be directly extracted from it. The field information for the two datasets is presented in Tab. III.

B. *Evaluation Metrics*

1) Keyword Evaluation Metric——F1: To assess the performance of keyword extraction methods, we use the F1 score

TABLE III

FIELD INFORMATION OF THE DATASETS

Retrieval Dataset		Test Set	
Invention Information	Abstract	Invention Information	Abstract
	Claims		Claims
	Description and description of drawings		Description
Metadata	Title		Description of drawings
	Kind Code	Metadata	Title
	Patent Type		Application number
	Inventors, Applicants		Publication number
	Filing Date	Retrieval Information	Keyword-based retrievable elements
	Application number		Related documents

to evaluate the top 5, 10, and 15 ranked candidates obtained from each method. Given the variability in the granularity of the results produced by different methods, we categorize the methods based on the average length of the extracted keywords:

Word-based Extraction Methods: Methods where the average length of extracted keywords is less than or equal to 3 characters. For example, TF-IDF typically extracts keywords with an average length of 2.16 characters.

Phrase-based Extraction Methods: Methods where the average length of extracted keywords is greater than 3 characters. For instance, phrase-based TextRank can produce keywords with an average length of up to 4.60 characters.

To ensure a fair evaluation, we reference the average length of the keywords in the gold standard(in Section IV-A2), which is 2.23 characters. When calculating the F1 score, duplicate candidates will be removed, and partial matching will be used. This means a keyword or phrase in the extracted results is considered correct if it fully or partially matches a keyword or phrase in the gold standard.

2) Evaluation Metrics for Retrieval Results: In practical scenarios, patent retrieval emphasizes recall; however, high recall rates necessitate greater review efforts. The Patent Retrieval Evaluation Score (PRES) [11] is specifically proposed for tasks such as patent retrieval, which prioritize recall. PRES focuses on the overall recall of the system and the user's review workload, estimating the quality of retrieval results based on the ranking of relevant documents.

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{max}} \quad (1)$$

Here, r_i denotes the rank of the $i-th$ relevant patent document in the retrieval results, n is the number of relevant patents in the collection, and N_{max} is the maximum number of documents a user is willing to review. PRES represents a balance between the recall rate of the retrieval results and the user's review effort.

C. Experiments

The experiment is divided into two phases, as illustrated in Fig. 2. First, keyword extraction methods are used to extract retrievable elements from the test set, and the F1 score relative to the gold standard(in Section IV-A2) is calculated to compare the performance of different methods. Second, the extraction results for each patent in the test set are combined into search queries, and the in-house developed retrieval engine JSS is employed for retrieval. The PRES value is calculated based on the recall rate of relevant documents in the retrieval results and their ranking, thereby evaluating the quality of the retrievable elements.

Given the lengthy and complex nature of patent application texts, we analyzed the proportion of retrievable elements in the test set that appear in the claims section to save time and financial resources associated with computation. Results indicated that retrievable elements appearing in the original application text have a probability of over 92% of appearing in the claims section. This aligns with the retrieval principles introduced in Section II-A regarding patent examination. Therefore, our experiment focuses on the ”claims” section of the application texts.

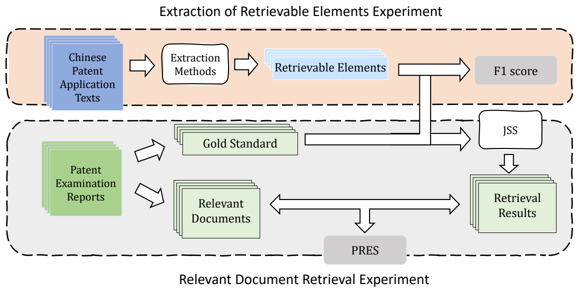
Extraction of Retrievable Elements Experiment
Chinese Patent Application Texts
Extraction Methods
Retrievable Elements
F1 score
Gold Standard
JSS
Patent Examination Reports
Relevant Documents
Retrieval Results
PRES
Relevant Document Retrieval Experiment

Fig. 2. Illustration of extraction of retrievable elements experiment and relevant document retrieval experiment.

1) Extraction of Retrievable Elements Experiment: We utilized the following methods to extract retrievable elements:

TF-IDF This method leverages the term frequency-inverse document frequency statistic. It extracts elements at the ”word” level, ranking candidate words based on their TF-IDF scores to produce the final extraction results.

Noun Phrase Frequency (Np Frequency) According to [6], extracting noun phrases from patents is more effective than extracting individual words. While the TF-IDF method does not perform as well as frequency-based methods, in this experiment, we implemented a noun phrase extraction method that ranks solely based on term frequency. This approach allows us to compare the performance of word-based versus phrase-based extraction in the context of patent retrieval.

TextRank This graph-based algorithm identifies the relationships between words or phrases through co-occurrence. The importance of nodes within this graph is used to rank and extract key elements. We conducted two experiments with TextRank: one at the ”word” level (TextRank-words) and another at the ”phrase” level (TextRank-phrases), using regular expressions on part-of-speech tagged text to achieve these granularity levels.

KeyBert and PromptRank Both KeyBert and PromptRank are methods based on pre-trained language models (PLMs). KeyBert utilizes PLMs to obtain embeddings for both documents and candidate words. It then ranks candidate words using cosine similarity. In our experiment, we employed the paraphrase-multilingual-MiniLM [14], trained on multilingual corpora, which we refer to as KeyBert-MiniLM.

PromptRank, on the other hand, connects text and candidates using prompts. It employs an Encoder-Decoder architecture to calculate the probability of candidate generation. Additionally, it applies position penalties based on the first appearance of candidates in the text. The combination of these factors is used to rank candidates. In our experiment, we encoded the template ”文本：[Document]” in the Encoder and ”这段文本是关于[Candidate]” in the Decoder. We calculated the probability of generating ”Candidate.”

PromptRank employs regular expressions to extract results at two granularities: ”words” and ”phrases.” We conducted experiments using the mengzi-t5-base-mt [20] and Randeng-T5-784M [18], trained on Chinese corpora, referred to as PromptRank-MZ-words, PromptRank-RD-words, PromptRank-MZ-phrases, and PromptRank-RD-phrases, respectively.

LLM Few-shot This approach utilizes the capabilities of large language models (LLMs) to extract search elements directly from the input content using few-shot learning, without requiring any parameter training. Inspired by the methodology outlined in Section II-C for extracting retrievable elements in patent retrieval, we employed the gpt-3.5-turbo-16k model with a prompt to extract retrievable elements effectively.

LLM Efficient-tuning This represents the sole supervised method utilized in our experiment. Given the scarcity of annotated data, efficient-tuning of large language models enables the training of an effective model for retrievable element extraction with minimal data. Following efficient-tuning on the Baichuan2-7b-chat model using a dataset comprising 1000 patent texts and their corresponding elements, we input the claims section of the test set into the efficiently-tuned model to obtain extraction results.

It is important to note that, except for LLM Few-shot and Efficient-tuning, other methods require initial text segmentation and part-of-speech tagging. Unless explicitly stated, stop words, function words, and other non-substantive terms are excluded. All remaining terms, regardless of their part of speech or type, are considered candidates for subsequent calculations.

2) Relevant Document Retrieval Experiment: This section introduces the retrieval engine used in the experiments and the experimental process.

Json Struct Search (JSS) Retrieval Engine The JSS retrieval engine, developed in-house, integrates various search modes, including symbolic retrieval, Boolean retrieval, and semantic retrieval. It supports the retrieval of structured and semi-structured documents like patent texts. JSS allows for the definition of custom search modes, search structures, and search content, supporting multiple forms of search expressions such as SQL queries, keywords, and semantic searches. For our experiments, JSS was used as the primary search engine.

Retrieval Experiment We indexed the entire content of the retrieval dataset using the JSS engine, creating a table named ”patent.” To minimize the influence of different search modes and expressions on the retrieval results, we employed the simplest symbolic retrieval mode. The top 10 extraction results were concatenated with commas to SQL for retrieval. An example of such a query is as follows:

SELECT TOP 50 id, title FROM patent WHERE fulltext LIKE ’终端, 云台, 支架’

This query indicates that the ”patent” table is being searched for patents whose full text (including titles, abstracts, claims and descriptions) contains the keywords ”终端”, ”云台” and ”支架”. It returns the top 50 most relevant patent application numbers (id) and their titles. In the retrieval experiment, the returned 50 patents are considered the retrieval results for the queried patents in the database, with N_{max} set to 50. The recall and the Patent Retrieval Evaluation Score (PRES) are calculated based on the number and ranking of the relevant documents retrieved. The gold standard retrieval results serve as a reference to evaluate the performance of different methods in the retrieval experiments.

D. Overall Results

We calculated the average length of the extraction results for different methods, as shown in Tab. IV. Based on the average length, all results were categorized into ”word-based extraction” and ”phrase-based extraction.”

In the first experiment, consider the retrievable elements provided by the examiner in each patent examination report as the gold standard.For each category, we computed the F1@5, F1@10, and F1@15 scores against the gold standard, as presented in Tab. V.

TABLE IV
AVERAGE LENGTH OF EXTRACTION RESULTS FOR GOLD STANDARD AND DIFFERENT METHODS

Data or Methods		Average length	Data or Methods		Average length
Gold Standard		2.23	Phrase-based	TextRank-phrases	4.6
Word-based	TF-IDF	2.16			
	Np Frequency	2.85		PromptRank-MZ-phrases	4.03
	TextRank-words	2.14			
	PromptRank-RD-words	2.43		PromptRank-RD-phrases	3.81
	PromptRank-MZ-words	2.32			
	LLM Efficient-tuning	2.33		LLM Few-shot	3.81
	KeyBert-MiNiLM	2.42			

The results indicate that among the word-based methods, the efficient-tuning method using large language models yielded effective results with minimal data, surpassing other unsupervised methods. For the remaining unsupervised methods, statistical methods like TF-IDF and Np Frequency performed relatively well. This suggests that in the absence of large datasets for model parameter tuning, statistical methods remain a viable option for extracting key information from patent texts.

In the phrase-based methods, the PromptRank method demonstrated superior performance. This is because PromptRank uses an Encoder-Decoder model to compute the generation probability of candidates, taking into account both the semantic information of the input text and the position of candidates within the original text. Moreover, in practical applications, phrase-level extraction methods are more suitable for automating the extraction of retrievable elements to aid in patent retrieval and examination. Compared to word-level results, phrase-level extractions provide more information and accuracy, making them more useful for professionals when evaluating, selecting, and refining automated results.

TABLE V
PERFORMANCE OF DIFFERENT METHODS ON THE TEST SET*

Methods		F1@5	F1@10	F1@15
Word-based	TF-IDF	0.243	0.240	0.217
	Np Frequency	0.197	0.249	0.260
	TextRank-words	0.204	0.210	0.203
	PromptRank-RD-words	0.217	0.240	0.234
	PromptRank-MZ-words	0.192	0.214	0.215
	KeyBert-MiNiLM	0.218	0.233	0.225
	LLM efficient-tuning	**0.330**	**0.342**	**0.342**
	KeyBert-MiNiLM	0.218	0.233	0.225
Phrase-based	TextRank-phrases	0.262	0.322	0.337
	PromptRank-MZ-phrases	**0.341**	**0.389**	**0.380**
	PromptRank-RD-phrases	0.340	0.379	0.371
	LLM few-shot	0.292	0.333	0.332

* The best results are bloded, the second-best results are underlined.

In relevant document retrieval experiment, the actual relevant documents provided in the examination reports were used as the ground truth. we calculated the recall R@10, R@30 and R@50, as well as the Patent Retrieval Evaluation Score (PRES), for each method's retrieval results, with N_{max} set to 50. The results are presented in Tab. VI. The retrieval results obtained using different search engines can vary. Therefore, we also employed the gold standard in the retrieval experiments, using its results in the JSS search engine as a reference. We used the gold standard retrieval results as a benchmark to assess the performance of various methods.

Interestingly, the retrievable elements generated by the LLM few-shot method outperformed the gold standard in terms of retrieval effectiveness. This suggests that the principles outlined in Section II-C—extracting search elements from the technical solutions in claims, focusing on aspects such as technical fields, problems, means, or effects—are essential. Utilizing computer algorithms to assist in this extraction process is feasible. If we can automate this principle-based extraction process, we could achieve similar or even better retrieval results.

Overall, phrase-based extraction results showed better performance in retrieval experiments compared to word-based extraction results. Comparing the results in Tab. V and Tab. VI, it is evident that phrase-based extractions that align more closely with the gold standard tend to have higher F1 scores. Although the LLM few-shot method's extraction results did not have outstanding F1 scores in Tab. V, likely due to different extraction processes, their retrieval performance was superior. This correlation was less pronounced for word-based extraction results.

The quality of retrievable elements is evaluated from two distinct stages of patent retrieval using the F1 score and PRES. The primary objective of patent retrieval is to find prior art that is most relevant to the current text, with the extraction of retrievable elements serving as an auxiliary means to facilitate the search process. The gold standard provided by professional examiners is a reference point but not the sole or optimal result. Consequently, the retrieval experimentsIV-C2 are more representative of the actual patent retrieval process, and the PRES metric can more accurately assess the effectiveness of search elements in practical retrieval scenarios.

TABLE VI
PERFORMANCE OF DIFFERENT METHODS IN RETRIEVAL EXPERIMENT*

Methods		R@10	R@30	R@50	PRES
Phrase-based	LLM few-shot	**0.598**	**0.743**	**0.791**	**0.706**
	PromptRank-RD-phrases	0.552	0.690	0.754	0.646
	gold standard	0.553	0.692	0.740	0.636
	PromptRank-MZ-phrases	0.530	0.675	0.732	0.631
	TextRank-phrases	0.371	0.499	0.558	0.462
Word-based	**gold standard**	0.553	0.692	0.740	0.636
	TF-IDF	**0.521**	**0.671**	**0.725**	**0.630**
	Np Frequency	0.505	0.650	0.701	0.610
	TextRank-words	0.488	0.634	0.687	0.603
	PromptRank-RD-words	0.480	0.628	0.691	0.596
	LLM efficient-tuning	0.481	0.628	0.689	0.577
	KeyBert-MiNiLM	0.453	0.604	0.671	0.564
	PromptRank-MZ-words	0.429	0.577	0.637	0.539

* The best results are bloded, the second-best results are underlined.

It is important to note that the experimental results, particularly those from the retrieval experiment, do not represent the upper limit of performance achievable by these methods on patent texts. This is because, up to this point in our work, we have only attempted the "preliminary retrieval" stage of the entire patent retrieval process mentioned in Section II-C. In practical retrieval scenarios, it is essential to continually adjust the retrievable elements and query expressions based on the retrieval results and expected outcomes. This includes expanding retrievable elements and expressions in terms of form, meaning, and perspective. Grouping retrievable elements to form different search expressions for retrieval is also important. Through multiple adjustments, more accurate and comprehensive retrieval results can be achieved.

V. CONCLUSION

In this paper, we present an overview of the general process of patent retrieval and explore the feasibility of using NLP techniques to assist in extracting retrievable elements, thereby enhancing patent retrieval and examination. We compare the effectiveness of different keyword extraction methods for patent retrieval tasks, including retrievable element extraction tasks and applying these extraction results to patent retrieval tasks. The experimental results indicate that both phrase-based keyword extraction methods and few-shot methods based on large language models are effective solutions for this task.

Patent retrieval and examination are exceedingly complex processes that require iterative cycles of "search, assessment, correction, and re-search". Each step relies on the substantial expertise of the retrieval staff, and the execution and outcomes of each step must be controllable and interpretable. This complexity reveals the limitations of general-purpose models or tools and highlights the necessity of domain-specific models or tools tailored for patents. Patent retrieval and examination cannot be entirely automated; they require human supervision and control. Thus, exploring interactive human-computer methods for retrieval and examination is a worthwhile research direction.

REFERENCES

[1] China National Intellectual Property Administration. *Patent Examination Guidelines in Chinese*. Intellectual Property Publishing House, 2023.

[2] Adrien Bougouin, Florian Boudin, and Béatrice Daille. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551, 2013.

[3] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. Yake! collection-independent automatic keyword extractor. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 806–810. Springer, 2018.

[4] Yao Cao and Cunfang He. An overview of determining basic search elements in chinese. *China Invention and Patent*, 18(Supplement):151–155, 2021.

[5] Liang Chen, Lili Chen, Haiyun Xu, Chao Wei, Na Su, and Weijiao Shang. Progress and prospects of patent mining research domestically and internationally in chinese. *Library and Information Service*, 68(2):110–133, 2024.

[6] Xu Chen, Zhiyong Peng, and Bin Liu. A summary of patent retrieval and analysis in chinese. *Engineering Journal of Wuhan University*, 47(3):420–425, 2014.

[7] Corina Florescu and Cornelia Caragea. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1105–1115, 2017.

[8] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020.

[9] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xiaoyan Bai. Promptrank: unsupervised keyphrase extraction using prompt. *arXiv preprint arXiv:2305.04490*, 2023.

[10] Bin Liu, Ling Feng, Fei Wang, and Zhiyong Peng. Patent retrieval and analysis to support technological innovation in chinese. *Journal on Communications*, 37(3):79–89, 2016.

[11] Walid Magdy and Gareth JF Jones. Pres: a score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 611–618, 2010.

[12] Debanjan Mahata, John Kuriakose, Rajiv Shah, and Roger Zimmermann. Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 634–639, 2018.

[13] CNIPA Patent Examination Cooperation (Jiangsu) Center of the Patent Office. *Learn Patent Retrieval with Examiners: A Quick Guide to Patent Information Retrieval in Chinese*. Intellectual Property Publishing House, 2019.

[14] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[15] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20, 2010.

[16] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 60(5):493–502, 2004.

[17] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860, 2008.

[18] Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970, 2022.

[19] Dongmei Xiao. *Intellectual Property Information Retrieval and Utilization in Chinese*. China Renmin University Press, 2021.

[20] Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*, 2021.