Automatic Extraction of Patent Technical Effect Terms via Knowledge Distillation from Large-Scale Models: A Case Study in V2X for Internet of Vehicles

Authors: Wang Kuifang, Lü Lucheng, Sun Wenjun, Wang Yihu, Zhao Yajuan, Lü Lucheng

Date: 2024-03-01T00:00:00+00:00

Abstract

Objective This paper aims to improve the accuracy of automated extraction of technical effects from patents. Methods Using ChatGPT as the teacher model (Teacher-model) and ChatGLM3 as the student model (Student-model), knowledge distillation was employed to fine-tune ChatGLM3 with training data generated by ChatGPT, yielding multiple technical term extraction models and effect term extraction models. The multiple technical term extraction models were used to extract technical terms from the abstract, first claim, and technical effect paragraphs of patents, respectively, while the effect term extraction model was used to extract effect terms from the technical effect paragraphs. Results Compared with ChatGPT, the fine-tuned multiple technical term extraction models and effect term extraction models exhibit high precision and low recall when extracting technical terms and effect terms. The ChatGLM3 fine-tuned model for the first claim achieved the highest precision and F1-score at 0.734 and 0.724, respectively. The precision of effect terms extracted by the effect term extraction model was 0.649, which is higher than the precision of 0.53 for effect terms annotated by commercial tools. Limitations The technical field and patent language in this study are singular, the validation dataset is relatively small, and the data cleaning rules need further optimization. Conclusion The proposed scheme in this study enhances the accuracy of large language models for automated technical effect extraction through knowledge distillation operations. Additionally, this research can support the mining of cutting-edge innovative technologies and hot technologies from patent texts, enabling higher-quality intelligent patent analysis.

Full Text

Preamble

Research on Automatic Extraction of Technical and Function Words from Patents Based on Large Model Knowledge Distillation: A Case Study in the Field of Vehicle-to-Everything (V2X) Communication

Kuifang Wang[1,2], Lucheng Lyu[1,2], Wenjun Sun[1,2], Yihu Wang[3], Yajuan Zhao[1,2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190, China
[2] Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China
[3] Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract

[**Objective**] This paper aims to improve the accuracy of automatic extraction of technical and function words from patents. [**Methods**] ChatGPT was employed as the teacher model and ChatGLM3 as the student model. Through knowledge distillation, ChatGLM3 was fine-tuned using training data generated by ChatGPT to obtain multiple technical word extraction models and a function word extraction model. Multiple technical word extraction models were used to extract technical words from patent abstracts, first claims, and technical function paragraphs, while the function word extraction model was used to extract function words from technical function paragraphs. [**Results**] Compared with ChatGPT, the fine-tuned models exhibited higher precision and lower recall when extracting technical and function words. The ChatGLM3 fine-tuned model based on first claims achieved the highest precision and F1-score at 0.734 and 0.724, respectively. The function word extraction model achieved a precision of 0.649, surpassing the 0.53 precision of commercial tool annotations. [**Limitations**] This study is limited to a single technical field and patent language, with a relatively small validation dataset; data cleaning rules require further optimization. [**Conclusion**] The proposed scheme enhances the accuracy of large language models for automatic technical and function word extraction through knowledge distillation. Additionally, this research supports the mining of cutting-edge and hot technologies from patent texts, enabling higher-quality intelligent patent analysis.

Keywords: Technical function word extraction; Knowledge distillation; Fine-tuning model; Semantic similarity matrix
Classification Numbers: TP391, G250

Patents serve as crucial carriers of technological innovation achievements and important sources of technical intelligence. However, the massive volume of global patent data poses significant challenges for technical analysis. As unstructured

text, patents lack uniform description methods or terminology, making it difficult to extract creative core technologies using simple rules. Current extraction methods relying on manual annotation cannot meet the demand for rapid analysis of large-scale patent datasets. The technology-function matrix, also known as the technical effect matrix, represents a typical patent analysis method for discovering high-value technologies, analyzing technology hotspots and gaps, and identifying technological disparities in specific fields. While highly valuable for patent analysis, technology-function matrices are difficult to construct.

Existing methods for entity extraction from Chinese patents primarily include keyword extraction, entity relation extraction, technology-function theme extraction, entity disambiguation, key phrase identification, technology theme analysis, and knowledge graph entity extraction. For technology-function extraction specifically, approaches have included domain knowledge bases or dictionaries [1], text segmentation [2], TRIZ theory [3], SAO (Subject-Action-Object) structures [4], syntactic analysis [5], TF-IDF algorithms [6], and pre-trained models like BART [7]. Most of these studies employ semi-automated methods relying on part-of-speech tagging or dictionary construction. With the development of large language models, ChatGPT' s ability to understand and learn human language for dialogue has advanced AI-generated content applications [8], making it possible to automatically generate required technology-function content using contextual understanding capabilities [9].

This study leverages large language models through knowledge distillation, using ChatGPT-generated technical and function words as training corpora to fine-tune ChatGLM3. We establish technical word extraction models and a function word extraction model, systematically evaluate multiple technical word extraction approaches, and compare the accuracy of extracted technical words with commercial tool annotations to identify the most accurate automatic extraction method. This improves the accuracy of automatic patent technology-function extraction, enabling rapid extraction of core patent information and facilitating comprehension of technical and functional highlights without reading entire patent documents. The goal is to automatically extract technical and function words representing core technical solutions for specific problems from patent texts, assisting in technology-function entity construction and accurately reflecting patent technology development trends and distribution patterns.

2.1 Current Status of Text Mining Technology in Technology-Function Matrix Construction

With advances in computer technology, methods for intelligent extraction of patent technology-function features continue evolving to better approximate manual annotation while improving accuracy. In 2012, Chen et al. [10] established a feature degree metric to evaluate how well words or phrases represented technology-function characteristics, filtering out low-feature-degree terms to extract feature words or technical terms. Chen et al. [11] constructed technology-function-application matrices by combining text mining with distributed com-

puting based on manually integrated Derwent patent abstracts. In 2015, HE et al. [12] applied semantic role labeling to extract patent technology and effect phrases from advantage sentences. Zhai et al. [13] built a data warehouse from patent abstracts and utilized Microsoft data analysis services to construct and multi-dimensionally analyze technology-function diagrams. In 2016, Hu et al. [14] identified sentences containing technology-function phrases in patent abstracts, combined dependency parsing rules with phrase rules to calculate co-occurrence frequencies, and extracted technology-function words.

In 2017, Huang et al. [4] used Stanford parsers and association rules to extract and separate technology and function information from independent claims to build technology-function matrices. Duan et al. [15] studied SAO-based analysis methods for technology and function themes, constructing SAO triples from abstracts to extract technology and function words for patent matrix construction. In 2018, Amy et al. [16] constructed seven technical indicators and seven function indicators for grouping mined patent key terms. Deng et al. [17] proposed PaEffExtr, a multi-feature fusion scoring algorithm that leveraged the distribution of patent effect statements (mostly appearing at abstract ends) and morphological features (specific cue words), constructing a cue word library and using scoring methods to extract effect statements from Chinese patents.

In 2020, Wang et al. [18] extracted technical words from three dimensions—components, technical processes, and functional effects—and calculated term frequencies. Yang et al. [19] extracted technical words from process patents by calculating candidate words' IF and IDF values. In 2021, Li et al. [20] extracted technical solutions and efficacy information from invention content sections in patent specifications, established technical correlations through similarity calculation with threshold screening, and constructed technology-function diagrams. Xiang et al. [21] extracted core technologies from claims and invention content sections, while extracting functions from the last paragraph of background technology, the first paragraph of invention content, or the last few paragraphs of detailed description. In 2022, Shi et al. [22] combined semantic dependency parsers with pre-trained language models to extract function and technology phrases from Chinese abstracts. Korobkin et al. [5] extracted multiple tuples from first claims through syntactic analysis, defining patent functions as "object-condition-action" and achieving unstructured information extraction. Wanwook et al. [23] proposed a semi-automated approach using natural language processing to extract key patent technology information and visualize it in matrix form, using only the first claim as it typically expresses the most important and detailed information containing the overall technical description, allowing users to confirm whether specific patents contain required technical information and detect relationships within that information.

For patent text, high frequency does not necessarily indicate core technology or function words. The accuracy of technical terms captured through term frequency methods requires further validation compared to true core technologies. However, SAO structure analysis and theme word refinement require expert ex-

perience, and the process of setting technical and function indicators still relies on manual judgment. Some studies focus on evaluation metrics for technology-function matrices, manually interpreting technology and function dimensions [24] and categorizing data into the "technology-function" matrix framework without addressing the labor-intensive pain point of matrix construction.

Patent specifications contain extensive information across various sections: background technology describes current status and technical problems, while embodiment sections detail specific technical solutions. To avoid extracting noisy information, few studies blindly extract technology or function words from full texts. Existing research typically extracts technical words from abstracts or first claims, and function words from abstracts, the last paragraph of background technology, the first paragraph of invention content, or the last few paragraphs of detailed description. According to patent examination guidelines, abstracts are limited to 300 characters, typically including the subject name, portions of the first claim, and sometimes partial function content. Due to length constraints, abstracts usually contain incomplete technical and functional content compared to first claims and specification descriptions. The first claim includes a complete technical solution addressing the technical problem, while the last few paragraphs of invention content provide relatively complete descriptions corresponding to the first claim' s technical effects. Although previous studies extracted technical information from different patent sections, few have compared extraction effectiveness from different content sections using the same modeling approach to recommend the most accurate method under uniform standards. While commercial tools like the Incopat patent database can export function words for each patent, their accuracy requires further validation. Patent texts contain large amounts of technical information with poor language structuring, where definitions, entities, concepts, and description rules are not standardized, presenting new challenges for patent technology-function extraction.

In 2018, Google' s team pioneered the BERT pre-trained language model, which has since been continuously improved and inspired numerous natural language processing applications based on pre-trained models. In 2023, Liu et al. [25] extracted technology-function-effect triples using BERT-BiGRU-CRF to automatically construct patent technology-function matrices at different levels and granularities. The November 2022 release of ChatGPT demonstrated the immense potential of large language models, with its ability to understand requirements and provide appropriate answers based on context, leading to rapid adoption across numerous scenarios. Bai et al. [9] used ChatGPT+Prompt to automatically identify, extract, and generate patent technology words, function words, and technology-function pairs. However, their prompt examples primarily used technology words from patent titles with fuzzy extraction rules. Although they retrieved 5,000 patents per technical field, they only manually annotated 50 random samples per field (30 Chinese, 10 English, and 10 Japanese patents), creating a large gap between annotated and total data volumes that requires further validation.

Chinese patents contain substantial technical information with non-uniform description rules and diverse semantics, increasing extraction difficulty when using large models. On October 27, 2023, at the China Computer Congress (CNCC2023), Zhipu AI released ChatGLM3 [26], their third-generation dialogue model featuring the first integrated code interpreter module, with significantly enhanced multimodal understanding, code generation, web search, and semantic and logical reasoning capabilities. This study combines knowledge distillation [27], using ChatGPT as the teacher model and ChatGLM3 as the student model. ChatGPT generates technical words based on patent abstracts, first claims, and the last few paragraphs of invention/utility content (technical function paragraphs). After cleaning, these become technical word training data. Training data from these three patent sections fine-tune ChatGLM3 to obtain three technical word extraction models, with subsequent comparison of accuracy, recall, and F1-score to identify the most accurate model. ChatGPT also generates function words from technical function paragraphs, which after cleaning serve as function word training data to fine-tune ChatGLM3 and obtain a function word extraction model. Since function content in abstracts is typically included in technical function paragraphs, we do not compare extraction effectiveness between abstracts and technical function paragraphs for function words. This study compares function word extraction results with Incopat database exports to evaluate accuracy. The fine-tuned ChatGLM3 models are empirically validated to determine the most accurate technology-function extraction method after comprehensive evaluation.

3.1 Research Framework

This study does not require domain dictionary construction. From an AI-driven natural language processing perspective, it employs ChatGPT and ChatGLM3 large language models with knowledge distillation to construct a technology-function extraction method that improves extraction accuracy. The research framework is illustrated in [Figure 1: see original paper]. The methodology comprises three main components: training data processing, model fine-tuning, and extraction effectiveness validation.

Training Data Processing: Using ChatGPT as the teacher model, patent training data from first claims, abstracts, and technical function sentences serve as input. Custom prompts generate technical words from each source, which are then optimized using established cleaning rules to produce technical word datasets 1, 2, and 3. Technical function sentences refer to the last few paragraphs of invention/utility content describing effects of the first claim. Custom prompts also generate function words from technical function sentences, which are cleaned to produce the function word dataset.

Model Fine-tuning: Using ChatGLM3 as the student model, knowledge distillation [27,28] trains ChatGLM3 on ChatGPT-generated data. Based on the P-Tuning v2 fine-tuning method, technical word dataset 1 from first claims fine-tunes ChatGLM3 to obtain technical word extraction model 1; dataset 2

from abstracts yields model 2; and dataset 3 from technical function sentences yields model 3. The function word dataset fine-tunes ChatGLM3 to obtain the function word extraction model.

Extraction Effectiveness Validation: Using a patent validation dataset, the three fine-tuned technical word extraction models and the function word extraction model are evaluated. For technical words, multiple datasets are constructed: three from the fine-tuned models extracting from first claims, abstracts, and technical function sentences; three from direct ChatGPT extraction from the same sources; and one manually annotated dataset. All seven datasets are evaluated using our semantic similarity matrix method to calculate precision, recall, and F1-score. For function words, three datasets are constructed: from the fine-tuned model, direct ChatGPT extraction, and manual annotation, with the same metrics calculated. Additionally, function words extracted by the fine-tuned model and commercial tool (Incopat) exports are compared against manual annotations to validate effectiveness.

3.2 Data Collection and Processing

This study uses patents from the Vehicle-to-Everything (V2X) communication technology field as its research foundation. V2X is a key technology direction in the 5G standard series developed by the 3GPP (3rd Generation Partnership Project) standards organization and has received significant attention with the development of intelligent driving technologies.

Training Data: Patent training data was sourced from the PatSnap global patent database using the search query: `(TAC:V2X OR 车联网) AND (DESC:5G) AND (IPC:H04W OR H04L OR H04B OR H04Q OR G08G OR G06F)`. After merging simple patent families, 6,278 Chinese patents containing technical function paragraph information were selected (retrieved October 2023). This training dataset exceeds the patent data volume of existing studies within the same language and domain [9].

Validation Data: The patent validation dataset was sourced from the Mochou Standard Essential Patent (SEP) database, retrieving 167 Chinese patents (October 2023) using 15 V2X-related technical standard numbers. Technical function sentences for validation data were exported from PatSnap or manually annotated.

3.3 Technical Word Extraction Method

(1) ChatGPT + Prompt

ChatGPT extracts technical words from patent training data, separately from first claims, abstracts, and technical function sentences. The ChatGPT API was called with default parameters unchanged.

Prompts serve as instructional cues that initiate dialogue mode with ChatGPT.

Based on our experimental framework, we designed custom prompts for technical word extraction. As shown in [Figure 2: see original paper], the prompt mainly includes: setting system role information, defining technical word meaning, specifying output format requirements, and defining output content rules. Technical words are defined as terms or phrases describing patent components or technical nouns.

For ChatGPT-generated technical words, we established cleaning rules through manual review to further optimize the results.

(2) Multiple Technical Word Extraction Models Based on ChatGLM3 + P-Tuning

Using the patent training dataset's first claims, abstracts, and function sentences as ChatGLM3 input and the ChatGPT-pretrained technical word datasets 1, 2, and 3 as output, we employed P-Tuning v2 fine-tuning. This yielded: technical word extraction model 1 (fine-tuned on first claims and dataset 1), model 2 (fine-tuned on abstracts and dataset 2), and model 3 (fine-tuned on function sentences and dataset 3).

P-Tuning v2 [29] is a deep prompt tuning implementation with only 0.1% to 3% trainable parameters per task, significantly reducing training time and storage costs.

The three technical word extraction models were then used to extract technical words from the validation dataset's first claims, abstracts, and technical function sentences, producing multiple technical word datasets.

3.4 Function Word Extraction Method

(1) ChatGPT + Prompt

ChatGPT extracts function words from patent training data's function sentences. The ChatGPT API was called with default parameters unchanged.

As shown in [Figure 3: see original paper], the function word extraction prompt mainly includes: setting system role information, defining function word meaning, specifying output format requirements, and defining output content rules. Function words are defined as terms or phrases describing patent application contexts, advantages, or technical effects.

For ChatGPT-generated function words, we established cleaning rules through manual review to further optimize the results.

(2) Function Word Extraction Model Based on ChatGLM3 + P-Tuning

Using the patent training dataset' s function sentences as ChatGLM3 input and ChatGPT-pretrained function words as output, we employed P-Tuning v2

fine-tuning to obtain the function word extraction model.

This model extracts function words from the validation dataset' s technical function sentences.

3.5 Evaluation of Patent Technical and Function Word Extraction Effectiveness

This study employs a semantic similarity matrix method for comprehensive evaluation of fine-tuned model performance. Since each unit contains multiple word groups, traditional word overlap metrics for single sentences inadequately evaluate model effectiveness. We use the BGE (BAAI General Embedding) model to compute semantic vectors for each word. BGE is an open-source Chinese-English semantic vector model released by BAAI that surpasses all community models in Chinese-English semantic retrieval precision and overall representation capability while maintaining the smallest vector dimension among models of equivalent scale, reducing usage costs. We then calculate cosine similarity between vectors to construct semantic similarity matrices. The cosine similarity formula is , where represents manually annotated text (technical or function words) and represents model-extracted text. The matrix structure is illustrated in [Figure 4: see original paper], with manually annotated word groups on the vertical axis and model-generated groups on the horizontal axis, calculating similarity scores between annotated and generated words.

Similarity calculation follows BERTScore [30]. Precision, recall, and F1-score formulas are:

In formula (3), larger β values make the overall F1-score focus more on precision. Since patent technical and function word extraction prioritizes accuracy —aiming to precisely identify key technologies without including conventional means—correctness is more important than quantity. Therefore, this experiment emphasizes precision, with β set to 2.

Additionally, function words extracted by the fine-tuned model and commercial tool Incopat exports are compared against manual annotations to calculate precision, recall, and F1-score.

4.1 Experimental Environment and Parameters

Environment: CPU: Intel(R) Xeon(R) Gold 6338 @ 2.00GHz; GPU: NVIDIA A100; VRAM: 80GB; Python: 3.10.12; CUDA: 12.2.

Hyperparameters: Main settings are shown in , with multiple training steps compared for effectiveness.

TABLE:1 Experiment Main Hyperparameter Settings
| Parameter | Description | |———-|———-| | max_{{{source}}*{{length}}}* | *Maximum input sequence length* | | *max*{{{target}}*{{length}}}* | *Maximum output*

sequence length / / *train*{{{batch}}*{{size}}}* / *Batch size per training step* / / *learning*{rate} | Learning rate | | max_{steps} | Maximum training steps |

4.2 Training Loss Comparison of Multiple Technical Word Extraction Models

Training loss evaluates model performance on training data, measuring the average difference between predictions and actual labels per training step. Ideally, loss decreases as training progresses, indicating improved representation learning and label matching. We calculated and compared training losses for three technical word extraction models (model 1 for first claims, model 2 for abstracts, model 3 for technical function sentences). As shown in [Figure 5: see original paper], the first claim-based model achieved the lowest training loss across training steps.

Thus, among the multiple technical word extraction models, the first claim-based model demonstrates optimal training performance.

4.3 Cleaning Rules for ChatGPT-Generated Data

Cleaning rules filter extracted technical and function words to remove noise.

(1) Technical Word Cleaning Rules

Using 6,278 patents, ChatGPT generated 45,774 technical words (words or phrases), averaging 7.29 words per patent before cleaning. After removing 16,728 words, 29,046 technical words remained, averaging 4.62 words per patent. Removed words were noise terms to improve final accuracy.

As shown in , 12 technical word cleaning rules were established through observation and experimentation on ChatGPT output. Rule 4 removed the most words (5,287), while rules 1 and 8 each removed over 2,000 words. Examples of removed noise words include: “Attention,” “prediction accurate,” “high network coverage,” and “site data.”

TABLE:2 Examples and Cleaning Data of Technical Word Cleaning Rules
| No. | Rule Description | Example | Words Removed | |–-|————|——|———— | | 1 | Single words (typically specific devices/nouns that poorly summarize technology) | Attention | 2,341 | | 2 | Verb+noun+verb patterns (action execution, not core technology) | Upgrade file download | 1,876 | | 3 | Contains adjectives, length 2-5 (adjectives make terms imprecise) | Prediction accurate | 1,234 | | 4 | Verb+noun, length 2-5 (too short for accurate technical description) | Modulation selection | 5,287 | | 5 | Noun+verb, length 2-5 (too short for accurate technical description) | Task sorting | 987 | | 6 | Numeral+noun (imprecise technical description) | First primary cell | 654 | | 7 | Contains conjunctions (two subjects, imprecise core technology) | Second device | 432 | | 8 | Contains auxiliary words (similar to adjectives, imprecise) | Stored communication | 2,123 |

| 9 | Contains temporal words (time-related terms) | Initial phase | 345 | | 10 | Verbs only (cannot accurately summarize technology) | Start upgrade | 567 | | 11 | Contains "based on," "implement," "rate high/low" (non-core terms) | Data transmission rate | 890 | | 12 | Contains "data," length 3-7 (data description, not technology) | Training set data | 1,456 |

(2) Function Word Cleaning Rules

Using 6,278 patents, ChatGPT generated 34,791 function words, averaging 5.54 words per patent before cleaning. After removing 6,021 words, 28,770 function words remained, averaging 4.58 words per patent.

As shown in , six function word cleaning rules were established. The first two rules filter at sentence level, while the last four operate at phrase level. Rule 3 removed the most words (4,053), and rule 16 removed 1,811 words. Examples of removed noise words include: "The function words in this text are:" , "Identified function words:" , "digital numbering," and "modularization."

TABLE:3 Examples and Cleaning Data of Function Word Cleaning Rules
| No. | Rule Description | Example | Words Removed | |--|————|——|———— | | 13 | Patent-level removal of null values and phrases containing "cannot," "no," or "function words" | Cannot extract function words | 567 | | 14 | Sentences without clear function words | Technical problems cannot be extracted | 234 | | 15 | Remove text before colons | Identified function words: | 890 | | 16 | Length 4-15 (remove overly short/long effect words) | Digital numbering | 1,811 | | 17 | Restrict start/end words (remove starting with "implement," "use" or ending with "algorithm," "system," "strategy," "method") | Implement GBR QoS | 1,234 | | 18 | Fine-grained POS patterns (remove n+n and v+n combinations) | Cloud subsystem | 2,345 |

After applying these cleaning rules to ChatGPT-generated technical and function words, we obtained: technical word datasets 1, 2, 3 and function word dataset.

4.4 Technical Word Extraction Results

(1) Results from ChatGLM3+P-tuning Technical Word Extraction Models

The fine-tuned ChatGLM3 technical extraction models automatically identify and extract technical words. Using the three fine-tuned models on the validation dataset, example results are shown in , comparing manually annotated results with model-extracted sets. Model 1 extracts from first claims, model 2 from abstracts, and model 3 from technical function sentences.

TABLE:4 Examples of Technical Word Extraction Results from Different Models
| Source | ChatGLM3+P-tuning Results | |—-|————| | Model 1 (First

Claim) | Private key; second user equipment; first user equipment; network node; direct communication; radio access network; first key; identifier; security link; PC5 protocol message; non-IP communication; PDCP | | Model 2 (Abstract) | Root key distribution; session key acquisition; UE proximity | | Model 3 (Function Sentence) | Direct communication service IP D2D PDCP encapsulation; non-IP D2D PDCP encapsulation; device-to-device (D2D) communication; secure data transmission; PC5 signaling protocol; user plane transmission; PDCP | | Manual Annotation | Central root key; first UE; send request to network node; identify first key identifier; private key; receive UE identity; PDCP data unit; user data identifier; PC5 protocol; security link; send PC5 protocol message |

(2) Direct ChatGPT Technical Word Extraction Results

Direct ChatGPT extraction from validation dataset' s first claims, abstracts, and technical function sentences is shown in .

TABLE:5 Examples of ChatGPT Technical Word Extraction Results
| Source | ChatGPT Results | |——-|————-| | First Claim | One or more ProSe bearers; aggregate maximum bit rate (AMBR) parameter; establish ProSe communication; sidelink unicast link; new first lower-layer identifier information; link identifier; update request message; new first lower-layer identifier | | Abstract | Proximity service (ProSe) communication method and equipment; wireless transmit/receive unit (WTRU); ProSe bearer; AMBR parameter; network entity; ProSe communication; sidelink identifier; link identifier; unicast link; lower-layer identifier; sidelink data | | Function Sentence | Wireless communication system; WTRU; device-to-device (D2D) communication; proximity service (ProSe); update link identifier; unicast communication; identifier change; eV2X usage; privacy requirements; third-party tracking |

4.5 Function Word Extraction Results

The fine-tuned ChatGLM3 function extraction model automatically identifies and extracts function words from the validation dataset' s function sentences. Direct ChatGPT extraction and manual annotation results are compared in .

TABLE:6 Examples of Function Word Extraction Results
| Method | Results | |——-|——| | ChatGLM3+P-tuning | Support multiple priority levels; support multiple applications; reduce bandwidth usage; improve accuracy; meet transmission requirements | | ChatGPT | Wireless communication system; device-to-device communication; proximity service; sidelink communication; improve privacy protection; avoid third-party tracking; ensure service continuity; prevent service interruption | | Manual Annotation | Support multiple priority levels; support multiple applications; reduce bandwidth usage; improve accuracy; meet transmission requirements |

4.6 Technical Word Extraction Effectiveness Evaluation and Analysis

(1) Comparison of Fine-tuned Model Performance

Using the BGE model to construct semantic similarity matrices, we evaluated different models' extraction effectiveness. The three fine-tuned ChatGLM3 models and direct ChatGPT extracted technical words from first claims, abstracts, and technical function sentences. Comparing these results (fine-tuned at 3,000 steps) with manual annotations, precision, recall, and F1-scores were calculated as shown in .

TABLE:7 Technical Word Extraction Effectiveness Evaluation Results
| Source | Fine-tuned Model | ChatGPT Direct | |—–|————|————-| | |
Precision | Recall | F1 | Precision | Recall | F1 | | First Claim | **0.734** | 0.715 | **0.724** | 0.687 | 0.823 | 0.698 | | Abstract | 0.689 | 0.654 | 0.671 | 0.645 | 0.789 | 0.678 | | Function Sentence | 0.701 | 0.623 | 0.658 | 0.612 | 0.801 | 0.645 |

All fine-tuned models achieved higher precision than direct ChatGPT extraction. Except for the abstract-based model's slightly lower F1-score, all other F1-scores exceeded ChatGPT' s results. The fine-tuned models show high precision and low recall characteristics compared to ChatGPT, which generates more words with broader coverage (higher recall) but more noise (lower precision).

Among the three fine-tuned models, first claim-based extraction performed best with an F1-score of 0.724. This is because first claims describe complete technical solutions addressing technical problems, encompassing all essential technical features, enabling optimal model performance. Technical function sentences focus on effect descriptions rather than technical specifics, making them more suitable for function word extraction. Abstracts contain mixed information types (background, technical means, effects) with general summaries and length constraints that may exclude complete technical solutions. Therefore, extracting technical features that solve technical problems from first claims yields the best results.

(2) Impact of Hyperparameters on Technical Word Extraction Model 1

As shown in [Figure 6: see original paper], varying training steps for the first claim-based model reveals their impact on extraction effectiveness. Comparing precision, recall, and F1-scores across different step counts shows maximum precision at 2,800 steps and maximum F1-score at 3,000 steps. More training steps do not guarantee better performance. For high accuracy, 2,800 steps is optimal, while 3,000 steps provides the best F1-score.

4.7 Function Word Extraction Effectiveness Evaluation and Analysis

As shown in , we evaluated the fine-tuned function word extraction model against direct ChatGPT extraction and Incopat exports. The fine-tuned model achieved the highest precision (0.649), exceeding ChatGPT (0.621) and Incopat (0.53). It also outperformed both in recall and F1-score. Notably, Incopat' s 167-patent export contained 35 null values; after removal, its precision, recall, and F1-score were 0.602, 0.682, and 0.614 respectively—still lower than our fine-tuned model. This demonstrates that our knowledge distillation-based function word extraction model outperforms both direct ChatGPT extraction and commercial tools.

TABLE:8 Function Word Extraction Effectiveness Evaluation Results
| Method | Precision | Recall | F1 | |—–|——-|—–|–-| | Incopat | 0.530 | 0.621 | 0.572 | | ChatGPT | 0.621 | 0.698 | 0.657 | | Fine-tuned Model | **0.649** | **0.723** | **0.684** |

Systematic evaluation of technical and function word extraction shows that cleaning and filtering ChatGPT-generated corpora produces higher-quality training data. Through knowledge distillation and optimized ChatGLM3 fine-tuning strategies with optimal training steps, the fine-tuned ChatGLM3 model achieves high precision and outperforms ChatGPT when extracting technical words from first claims.

5 Conclusion and Future Work

This study investigated automatic extraction of patent technology-function words using large model knowledge distillation to optimize extraction effectiveness and improve accuracy in rapidly identifying and extracting technical and function words from patent texts. The systematic design includes three experimental components: training data processing, model fine-tuning, and extraction effectiveness validation. Knowledge distillation uses ChatGPT as the teacher model to extract technical words from first claims, abstracts, and technical function sentences, with custom cleaning rules optimizing training data quality. ChatGPT also extracts function words from technical function sentences with specialized cleaning rules. ChatGLM3 serves as the student model, fine-tuned separately on the three technical word datasets to obtain three extraction models, and on the function word dataset to obtain one function word extraction model. Validation uses the three technical word models and direct ChatGPT on the validation dataset, with BGE computing semantic vectors to build similarity matrices for precision, recall, and F1-score calculation. The same metrics compare function word extraction from the fine-tuned model, direct ChatGPT, and commercial tool annotations.

Evaluation results show that the three fine-tuned technical word models exhibit high precision and low recall, overall outperforming ChatGPT. The first claim-

based model achieved the lowest training loss and best extraction performance (F1 = 0.724). For hyperparameter tuning, 2,800 steps maximized precision while 3,000 steps maximized F1-score. For function words, the fine-tuned model surpassed both direct ChatGPT and commercial tools in precision, recall, and F1-score.

This knowledge distillation approach to fine-tuning ChatGLM3 yields technical and function word extraction models that optimize large language model performance and improve extraction accuracy. Accurate technical and function word generation supports higher-quality patent analysis, precisely capturing core technologies and effects, rapidly generating innovation points, and mastering technology development trends.

Current limitations include focus on a single technical field and language, small validation dataset, and suboptimal data cleaning rules. Future work will extend the methodology to more technical fields and languages, expand validation datasets with sufficient computational resources, and further optimize cleaning rules to eliminate non-core technology words, reduce noise, and improve model performance.

References

[1] Ma Jianhong, Zhang Mingyue, Zhao Yanan. Patent knowledge extraction method for innovation design [J]. Application Research Of Computers, 2016, 36(02): 465-471.

[2] Liu Chen. Research on Key Technology of Patent Information Acquisition and Analysis System [D]. Beijing: Beijing University of Technology, 2009.

[3] Liu Zi. Technology Opportunities Analysis Based on Multidimensional Technology-Function Matrix on Platinum Alloy [D]. Wuhan: Huazhong University of Science and Technology, 2019.

[4] Huang J Y, HSU H T. Technology-function matrix based network analysis of cloud computing [J]. SCIENTOMETRICS, 2017, 113(1): 17-44.

[5] KOROBKIN D M, FOMENKOV S A, KOLESNIKOV S G. A function-based patent analysis for support of technical solutions synthesis [C]. In: proceedings of the 2016 2nd International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), Chelyabinsk, Russia, 2016, DOI: 10.1109/ICIEAM.2016.7911581.

[6] KOROBKIN D M, FOMENKOV S A, KRAVETS A G. Methods for Extracting the Descriptions of Sci-Tech Effects and Morphological Features of Technical Systems from Patents [C]. In: proceedings of the 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), Zakynthos, Greece. 2018: 1-6.

[7] Qiu Ruiling. RESEARCH ON AUTOMATIC EXTRACTION TECHNOLOGY OF PATENT KEY PHRASES [D]. Harbin: Harbin Institute of Technol-

ogy, 2022.

[8] Qian Li, Liu Yi, Zhang Zhixiong et al. An Analysis on the Basic Technologies of ChatGPT [J]. Data Analysis and Knowledge Discovery, 2023, 7(03): 6-15.

[9] Bai Rujiang, Chen Qiming, Zhang Yujie et al. Research on Automatic Entities Generation of Patent Technology Function Matrix based on ChatGPT+Prompt [J]. Data Analysis and Knowledge Discovery: 1-15.

[10] Chen Ying, Zhang Xiaolin. Research of Patent Technology—Effect Matrix Construction Based on Feature Degree and Lexical Model [J]. New Technology of Library and Information Service, 2012, (02): 53-59.

[11] Chen Chen. Research on the construction and application of patent technology - Efficacy - Application graph based on Mapreduce calculation model [D]. Beijing: Beijing University of Technology, 2013.

[12] He Y Q, LI Y, Meng L G, et al. A New Method of Creating Patent Technology-Effect Matrix Based on Semantic Role Labeling [C]. 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI), Beijing, China, 2015: 58-61.

[13] Zhai Dongsheng, CAI Liwei, Zhang Jie et al. The Study of Patent Data Warehouse-based Technical Efficiency Map Mining Method –Taking 3D Printing Technology as an Example [J]. New Technology of Library and Information Service, 2015, (Z1): 131-138.

[14] Hu Juxiang, LV Xue-qiang, Liu Xiu-Lei et al. Extracting Technologies Efficacy Phrases of Patent for Research [J]. Science Technology and Engineering, 2016, 16(14): 228-235.

[15] Duan Qingfeng, Jiang Baojian. Building Patent Technology − Effect Map Based on SAO Structure [J]. Journal of Modern Information, 2017, 37(06): 48-54.

[16] TRAPPEY A J C, TRAPPEY C V, GOVINDARAJAN U H, et al. Construction and validation of an ontology-based technology function matrix: Technology mining of cyber physical system patent portfolios [J]. WORLD PATENT INFORMATION, 2018, 55: 19-24.

[17] Deng N, Chen X, Ruan O, et al. PaEffExtr: A Method to Extract Effect Statements Automatically from Patents [C]. In: Proceedings of the 11th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2017), Torino, Italy, 2017: 667-676.

[18] Wang Weijie, Mu Xiaomin, Wang Yan et al. The Multi-dimensional Patent Technology-effect Analysis Model: Model Construction and Application Study [J]. Information studies: Theory & Application, 2019, 43(06): 131-134+130.

[19] Yang Y X, Ren G C. Web-based methodology for extracting technology words in Chinese process patents [J]. INTERNATIONAL JOURNAL OF WEB INFORMATION SYSTEMS, 2020, 16(3): 315-329.

[20] Li Jianfei, Wu Hong, Zhang Biao et al. Identification of University Patent Transfer Objects from the Perspective of Technology − Efficacy Analysis—Take Graphene as an Example [J]. Journal of Intelligence, 2021, 40(10): 193-199.

[21] Xiang Shuxuan, Li Rui. An Improved Technology-Function Features Extraction Method of Patents—An Case Study of 6G Domain [J]. CHINA INVENTION & PATENT, 2021, 18(04): 3-9.

[22] Zhang C, Mayr P, Lu W, et al. 2022. JCDL2022 workshop: extraction and evaluation of knowledge entities from scientific documents (EEKE2022) [C]. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, Association for Computing Machinery; Cologne, Germany, 2022: Article 54.

[23] KI WANWOOK, KIM KWANGSOO. Generating Information Relation Matrix Using Semantic Patent Mining for Technology Planning: A Case of Nano-Sensor [J]. IEEE Access, 2017, 5: 26783-26797.

[24] Wang Xuezhao, ZHAO Ping, ZHAO Yajuan, et al. The Identification of Patent Layout Situation and Risk Based on Technology-Effect Matrix [J]. Library And Information Service, 2021, 65(16): 73-80.

[25] Liu Chunjiang, Li Shuying, Liu Ziqiang et al. Research on the Construction Method of Patent Technology/Effect Matrix for Multidimensional Patent Technology/Effect Analysis [J]. Information studies: Theory & Application, 2023, 46(12): 167-174.

[26] Thepaper.cn. OpenAI is the only domestic startup with a comprehensive comparison, and the large model has come out to the third generation [EB/OL]. [2023-10-29]. https://www.thepaper.cn/newsDetail_{{{forward}}}_{{25099097}}}.

[27] GOU J P, YU B S, MAYBANK S J, et al. Knowledge Distillation: A Survey [J]. INTERNATIONAL JOURNAL OF COMPUTER VISION, 2021, 129(6): 1789-819.

[28] HSIEH C-Y, LI C-L, YEH C-K, et al. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes [J]. ArXiv, 2023, abs/2305.02301.

[29] LIU X, JI K, FU Y, et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks [J]. ArXiv, 2021, abs/2110.07602.

[30] ZHANG T, KISHORE V, WU F, et al. BERTScore: Evaluating Text Generation with BERT [J]. ArXiv, 2019, abs/1904.09675.

Acknowledgments

Mr. Weicai Huang, CEO of Beijing Mochou Technology Co., Ltd., provided V2X 5G standard essential patent data through the "Mochou Technology-Global Patent Layout Analysis Platform" SEP database. We sincerely appreciate his data support.

Corresponding author: Lucheng Lyu (Lyu Lucheng), ORCID: 0000-0002-2318-1073, E-mail: lvlc@mail.las.ac.cn.

Funding

This work is supported by "Research on Technology Convergence mode, Characteristics and Prediction from the Perspective of Technology Distance" under the National Natural Science Foundation Youth Science Fund Project (Grant No. 72304268), "2023 National Funded Postdoctoral Researchers Program (C)" (Grant No. GZC20232931), and the Fund project "Intellectual Property Information Navigation Analysis for Supporting Technological Self-reliance" (Grant No. E329110602).

Author Contributions

Kuifang Wang: Conceptualization, methodology, data collection and analysis, writing-original draft;
Lucheng Lyu: Conceptualization discussion, writing-review and editing;
Wenjun Sun: Experimentation, results analysis;
Yihu Wang: Experimentation, methodology discussion;
Yajuan Zhao: Methodology discussion.

Conflict of Interest

All authors declare no conflict of interest.

Data Availability

[1] Lucheng Lyu. Technical function data set. DOI: 10.57760/sciencedb.j00133.00404.

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