An Improved Meta-Knowledge Prompt Engineering Approach for Generating Research Questions in Scientific Literature

Meng Wang[1], Zhixiong Zhang[1,2], Hanyu Li[1,2] and Guangyin Zhang[2]
[1]*National Science Library, Chinese Academy of Science, Beijing, China*
[2]*University of Chinese Academy of Science, Beijing, China*

Keywords: Research Question Generation, Prompt Engineering, Knowledge Extraction, LLMs, Knowledge-Rich Regions.

Abstract: Research questions are crucial for the development of science, which are an important driving force for scientific evolution and progress. This study analyses the key meta knowledge required for generating research questions in scientific literature, including research objective and research method. To extract meta-knowledge, we obtained feature words of meta-knowledge from knowledge-enriched regions and embedded them into the DeBERTa (Decoding-enhanced BERT with disentangled attention) for training. Compared to existing models, our proposed approach demonstrates superior performance across all metrics, achieving improvements in F1 score of +9% over BERT (88% vs. 97%), +3% over BERT-CNN (94% vs. 97%), and +2% over DeBERTa (95% vs. 97%) for identifying meta-knowledge. And, we construct the prompts integrate meta-knowledge to fine tune LLMs. Compared to the baseline model, the LLMs fine-tuned using meta-knowledge prompt engineering achieves an average 88.6% F1 score in the research question generation task, with improvements of 8.4%. Overall, our approach can be applied to the research question generation in different domains. Additionally, by updating or replacing the meta-knowledge, the model can also serve as a theoretical foundation and model basis for the generation of different types of sentences.

1 INTRODUCTION

Research questions play a crucial role in revealing the specific content of scientific and technological literature and grasping the research theme of an article., which serve as both the logical starting point and the guiding core of scientific research (Kuhn, 1962). Scientific literature, as an essential medium for recording scientific knowledge, is essentially a record and description of the process of proposing and solving research questions. Research question sentences are a crucial component of the knowledge content in scientific literature. By identifying research question sentences in scientific literature, we can explore the knowledge content contained within. It can be said that grasping the research question sentences of an article is an important prerequisite for understanding the content of a piece of scientific literature. Therefore, it will be of great significance to automatically identifying or generating research questions in scientific literature.

However, there are two limitations to current researches about identifying or generating research questions. Firstly, most current studies are mainly based on training on general datasets, ignoring the meta knowledge required for specific domains or tasks. Secondly, even if domain data is used for training LLMs, they have not filtered and refined the meta knowledge in scientific literature, and still mix a lot of redundant information. Therefore, we attempt to propose a research question generation method based on meta-knowledge prompt engineering. To extract key meta knowledge required for generating research questions from scientific literature, a sentence classification model based on feature word vectors is proposed. Then, research question generation prompts that integrate meta-knowledge will be used to fine-tune LLMs, which will provide more accurate and targeted input, thereby improving the quality and accuracy of the generated results. The architecture of the proposed method in this paper is shown in Figure 1.

The main contributions of this paper are as follows:

(1) To improving the quality and accuracy of the generated results, the prompts that integrate meta-

Wang, M., Zhang, Z., Li, H. and Zhang, G.
An Improved Meta-Knowledge Prompt Engineering Approach for Generating Research Questions in Scientific Literature.
DOI: 10.5220/0013060900003838

In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 1: KDIR*, pages 457-464
ISBN: 978-989-758-716-0; ISSN: 2184-3228

knowledge are constructed and used to fine tune LLMs.

(2) To extract key meta knowledge required for generating research questions from scientific literature, an improved DeBERTa model considering the feature word vectors is proposed.

(3) To improve the efficiency of meta-knowledge extraction, sections and paragraphs containing meta knowledge are located in scientific literature.

(4) The constructed prompt dataset that integrates meta knowledge is used to fine-tune LLMs.

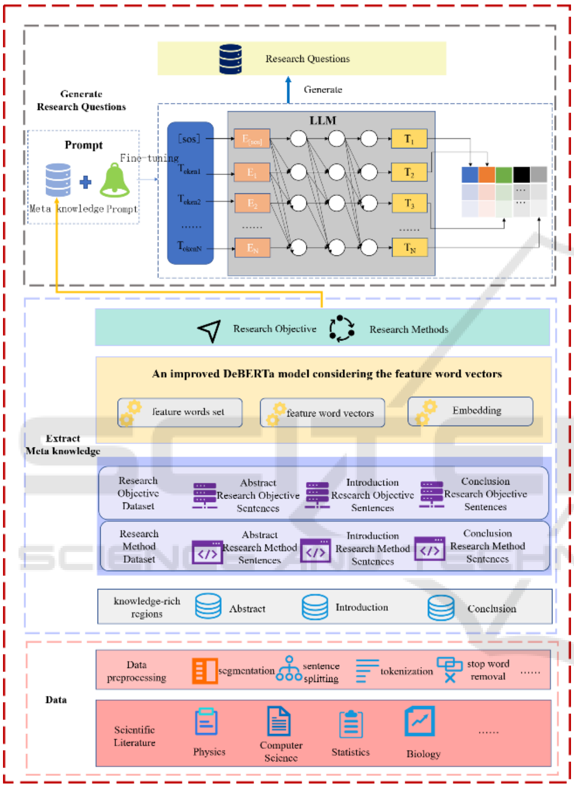

Figure 1: The architecture of the meta-knowledge prompt engineering approach for generating research questions in scientific literature.

The rest of this paper is organized as follows. The existing research of the meta-knowledge extraction and prompt engineering is presented in Section 2. Section 3 discusses an improved DeBERTa model, which considers the feature word vectors and knowledge-rich regions to extract key meta knowledge from scientific literature. The prompts that integrate meta-knowledge are constructed and used to fine tune LLMs in Section 4. Finally, Section 5 ends this study with conclusions and future work.

2 LITERATURE REVIEW

2.1 Meta-Knowledge Extraction

Meta-Knowledge extraction, also known as information extraction, refers to the task of automatically extracting structured information from unstructured or semi-structured text (Sarawagi, 2008). It aims to identify and extract relevant entities, relations, and events from text data, converting them into a structured format that can be easily processed and analyzed by downstream applications (Martinez-Rodriguez et al., 2018). Knowledge extraction plays a vital role in various natural language processing (NLP) applications, such as question answering, information retrieval, and knowledge graph construction (Chowdhary & Chowdhary, 2020).

In recent years, two mainstream approaches have emerged in the field of knowledge extraction: methods based on pre-trained models and methods based on LLMs. Methods based on pre-trained models utilize language models pre-trained on large-scale unlabeled text data, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020), and fine-tune them for specific knowledge extraction tasks. Chen et al. further explored the potential of DeBERTa for knowledge extraction by proposing a novel framework called DeBERTa-KE. This framework leverages the power of DeBERTa to jointly extract entities and relations from text, enabling end-to-end knowledge extraction (Chen et al., 2021).

With the growth of computational resources and the expansion of training data, large language models such as GPT (OpenAI, 2023), LLaMA (Touvron et al., 2023), and ChatGLM (Zeng et al., 2023) have demonstrated remarkable capabilities in the field of natural language processing. Researchers have begun to explore the use of these large language models for knowledge extraction tasks. The Meta AI team open-sourced the LLaMA model, which has 65 billion parameters. However, as the key information of sentences, feature words directly reflect the main content and deep meaning of the sentence and play an important role in improving the accuracy of research question sentence identification. Therefore, it is necessary to consider feature words in knowledge extraction (Touvron et al., 2023).

2.2 Prompt Engineering

Prompt engineering, also known as prompt design or prompt optimization, refers to the process of designing and optimizing prompts to effectively elicit desired behaviors or outputs from language models (Liu et al., 2023). It involves carefully crafting input prompts that guide the language model to generate high-quality, relevant, and coherent text. The quality of the generated text heavily depends on the effectiveness of the input prompts (Reynolds & McDonell, 2021). Well-designed prompts can significantly improve the coherence, relevance, and accuracy of the generated text, while poorly designed prompts can lead to nonsensical, irrelevant, or even harmful outputs.

We summarized the researches of meta-knowledge extraction and prompt engineer, and the results showed that there are two primary deficiencies in the current research: (1) many studies do not consider feature words in identifying research question sentences; (2) some researchers only use prompts to fine tune LLMs, which ignoring the meta knowledge required for specific domains or tasks. Therefore, this study analyzes the meta-knowledge required for generating research questions and manually summarizes the feature words of them. Moreover, the feature words are then embedded into the extraction model to improve the accuracy of the meta-knowledge extraction. The extracted meta-knowledge is integrated into prompt engineering to train LLMs, thereby enhancing the quality of the generated research questions.

3 EXTRACTING META KNOWLEDGE

3.1 Analysis of Meta-Knowledge Required for Research Question Generation

As the starting point and core of scientific research, research questions determine the direction, content, and objectives of a study. Generally, research questions can be divided into two main categories: theoretical questions and methodological questions (Alvesson & Sandberg, 2013). Theoretical questions focus on exploring the essence, laws, and mechanisms of things, aiming to establish or develop scientific theories. In scientific literature, these questions are usually reflected in the research objective section, where researchers explicitly state the theoretical issues they intend to explore.

Methodological questions arise from the challenges encountered in the technical methods during the research process, aiming to explore effective solutions. In scientific literature, methodological questions are usually reflected in the sentences about research method, where researchers focus on introducing the specific technical solutions and implementation steps adopted to solve the problems.

Thus, the key meta-knowledge required for generating research questions from scientific literature in this paper are the sentences of research objective and method, respectively.

3.2 Feature Word Vector Construction

3.2.1 Feature Word Sets

This paper employs manual annotation and iteration-based semi-automatic annotation methods to construct a dataset of research objective sentences and method sentences, obtaining a total of 20,000 high-quality corpus entries. From a linguistic perspective, feature words and characteristic sentence patterns in these two types of sentences are analyzed to construct a basic feature word set.

By combining the grammatical positions and contextual information of feature words, this paper obtains a total of 40 feature words. Some of the feature words and their contexts are shown in Table 1.

Table 1: Some feature words and the contexts.

feature words	contexts
analyze	…… were analyzed
	In order to analyze ……
	This paper analyzes ……
propose	…… was proposed in this study.
	In this paper …… is proposed
	This paper proposes ……
study	……was studied in this paper
	…… was studied

3.2.2 Feature Word Vector

Based on the analysis of part-of-speech tags and syntactic structure types of feature words, this paper calculates the frequency of feature words appearing in predicate positions and further expands the basic feature word set. A total of 40 feature words for knowledge elements are obtained, with a total frequency of 22,400 (notably, a sentence may contain multiple predicates). The proportion of each feature word represents its weight. Table 2 shows the frequency and weight distribution of some feature words.

Table 2: The frequency and weight distribution of some feature words.

feature words	frequency	weight
propose	5619	0.2508
explore	2520	0.1125
analyze	1955	0.0873
study	1702	0.0760
investigate	1549	0.0692
……	……	……
Total	22400	1

3.3 Embedding Feature Word Vector

This paper considers embedding the weight information of feature word vectors directly in the Classifier output stage within the DeBERTa model. The specific working mechanism of embedding feature word weight information is as follows:

Assume that for each input sentence, the DeBERTa model generates a hidden state vector $\mathcal{H} =< h_0, h_1, h_2, \ldots, h_L >$, where the dimension is L. In the DeBERTa model, the dimension of the hidden state vector is generally 768. The weight vectors of feature words constructed in this paper is $\mathrm{F} =< f_0, f_1, f_2, \ldots, f_{feature\_dim} >$, where f_n is the weight of the n^{th} feature vector. However, when the input sentence does not match any feature word, $f_n = 0$. $feature\_dim$ is the dimension of the feature vector. The hidden state vector of the RoBERTa model and the feature vector weight are concatenated, and this operation is implemented in the forward method of the Roberta Classifier, i.e.:

$$\mathcal{H}' = concat(h, f) \tag{1}$$

The dimension of the concatenated vector $\mathcal{H}'$ is L + feature_dim.

The linear layer of the classifier processes the concatenated vector $\mathcal{H}'$, and the formula is as follows:

$$\text{logits} = \mathrm{W} \cdot \mathcal{H}' + b \tag{2}$$

where W is the weight matrix with a dimension of L + feature_dim. b is the bias vector. logits is the raw score output by the classifier, which is finally passed to the softmax function to obtain the predicted probability distribution:

$$P = \text{softmax}(\mathrm{W} \cdot \mathcal{H}' + b) \tag{3}$$

3.4 Extracting Meta Knowledge

3.4.1 Locating Knowledge-Rich Regions

In scientific literature, knowledge is not evenly distributed but exhibits certain concentrations and regularities (Fortunato et al., 2018). Therefore, under the constraints of this writing logic, research objective sentences and research method sentences tend to be concentrated in the specific sections or paragraphs mentioned above. The knowledge-rich regions of research objective sentences and method sentences are shown in Figure 2.

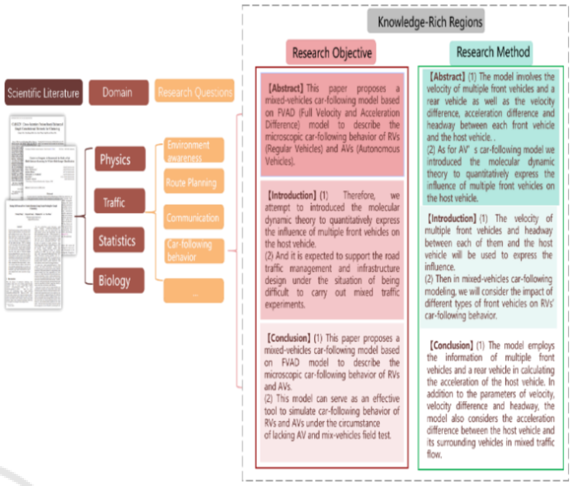

Figure 2: The knowledge-rich regions of research objective and method.

3.4.2 Experiment

This paper selects the full text of scientific literature and extracts the abstract, introduction, and conclusion sections by locating fine-grained knowledge-rich regions. The training corpus is divided into training, validation, and test sets according to the ratio of 8:1:1, ensuring the consistency of positive and negative sample distributions across the datasets. The dataset format is shown in Table 3.

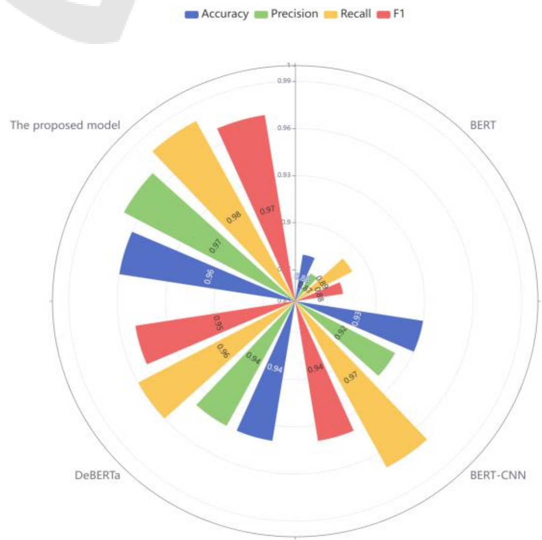

Figure 3: The extraction results of different models.

This paper selects BERT, BERT-CNN (Safaya et al., 2020), and DeBERTa as baseline models. According to Ref. (Li et al., 2023) and (Mei et al., 2023), the hyperparameter settings are shown in Table 4, and the extraction results are presented in Figure 3. The experimental results demonstrate that compared to the other three types of baseline models, the DeBERTa model based on feature word vectors proposed in this paper achieves the best meta-knowledge extraction performance, with an F1 score of 0.97.

Table 3: The dataset format.

Label	Sentence
0	Developing sharing economy of forestry has become an option to promote forestry development and solve the problems emerging from forestry economy.
1	In order to reveal the properties of polar metabolome in inflammatory cells, we selected LPS-induced RAW264.7 inflammatory cell models as the carrier for the research of metabolic fingerprint analysis.
2	As for AV's car-following model we introduced the molecular dynamic theory to quantitatively express the influence of multiple front vehicles on the host vehicle.

Table 4: The Hyperparameters of different models.

Hyperparameters	BERT	BERT-CNN	DeBERTa
lr	2e-5	2e-5	1e-5
b	16	16	64
e	30	30	10

4 GENERATING RESEARCH QUESTIONS BASED ON META-KNOWLEDGE PROMPT ENGINEERING

4.1 Constructing Meta-Knowledge Prompt Engineering

Considering the two key meta-knowledge elements of research questions: research objective sentences and research method sentences, we integrate the above-mentioned meta knowledge to manually construct the prompts, aiming to provide more accurate knowledge input for LLMs and improve the quality of research question generation. The format of the research question prompt is as follows: Given the title: "Title", the research objective: "Research Objective Sentence", the research methods: "Research Method Sentence", can we distill a concise question summarizing the research issue addressed in this article? Please use appropriate question words! Question: "Summarized Research Question". This prompt includes three knowledge elements: paper title, research objective sentence, and research method sentence, which are integrated into a complete research question generation task description, and finally provides a manually summarized research question.

This paper manually constructs 2,000 research question generation prompts, and some examples are as follows: Given the title: "Interpolating between Images with Diffusion Models", the research objective: "One little-explored frontier of image generation a........", the research methods: "We apply interpolation in the latent space \.......", can we distill a concise question summarizing the research issue addressed in this article? Please use appropriate question words! Question: How can we enable interpolation between two images using diffusion models, a capability missing from current image generation pipelines?

4.2 Fine-Tuning LLMs for Research Questions

To improve the quality of research question generation, this paper fine-tunes LLMs using the constructed prompt dataset that integrates meta knowledge. The fine-tuning dataset consists of three parts: task description, input, and output. The task description clearly states the objective of the generation task, what kind of task the model needs to complete, and what specific requirements and constraints exist, providing clear guidance for the subsequent input and output.

Based on the constructed fine-tuning dataset for research questions that integrates fine-grained knowledge, we adopt the LoRA (Low-Rank Adaptation) fine-tuning approach to fine-tune the large model (Su et al., 2021). The hyperparameter settings are as follows: batch_size: int = 10, micro_batch_size: int = 2, num_epochs: int = 2, learning_rate: float = 1e-5, lora_r: int = 8, lora_alpha: int = 16, lora_dropout: float = 0.05. The core idea of LoRA is to introduce a set of low-rank projection matrices at each layer of the large model and optimize these matrices to adapt the original model.

Specifically, for the i^{th} layer of the model, LoRA defines two projection matrices A_i and B_i with dimensions (d, r) and (r, d), respectively, where d is the hidden layer dimension of the model, r is the projection dimension, and $r \ll d$.

During forward propagation, LoRA adds a correction term based on the projection matrices to the original layer computation result. Suppose the original forward computation of the i-th layer can be represented as:

$$h_i = f_i(x_i) \quad (4)$$

where x_i is the input of the i^{th} layer, and f_i is the forward computation function of the i^{th} layer (such as self-attention, feed-forward network, etc.). In LoRA, the modified forward computation formula is:

$$h_i{}' = f_i(x_i) + A_i B_i f_i(x_i) = f_i(x_i) + \Delta_i \quad (5)$$

where $\Delta_i = A_i B_i f_i(x_i)$ represents the correction term introduced by LoRA. This correction term can be seen as adding a low-rank perturbation to the original layer output $f_i(x_i)$.

The optimization objective of LoRA is to minimize the loss function of the modified model on the new task:

$$\mathcal{L}(\theta, \{A_i, B_i\}_{i=1}^{L}) = \sum_{(x, y)\in\mathcal{D}} \ell\left(f_{\theta, \{A_i, B_i\}}(x), y\right) \quad (6)$$

where θ represents the fixed parameters of the original model, $\{A_i, B_i\}_{i=1}^{L}$ represents all the projection matrices introduced by LoRA, $\mathcal{D}$ is the training dataset of the new task, and ℓ is the task-related loss function (such as cross-entropy loss). During the optimization process, we only update $\{A_i, B_i\}_{i=1}^{L}$, while keeping θ unchanged. Therefore, the training overhead of LoRA is much smaller than that of traditional full-parameter fine-tuning. At the same time, since the rank *r* of the projection matrices is much smaller than the dimension *d* of the original model, the additional parameters introduced by LoRA are also much smaller than the original model. The fine-tuning experimental results of different models are shown in Table 5.

4.3 Experimental Results and Analysis

To verify the effectiveness of the research question generation method that integrates meta-knowledge extraction, this paper selects Mistral-7B (Devillers et al., 2023), Baichuan2-7B (Wu et al., 2023), Chatglm3-13B (Zeng et al., 2022), Internlm-7B (Cai et al., 2024), and Llama3-8B (Touvron et al., 2023) as benchmark models. We compare the quality of the generated research questions with and without fine-tuning, and use the Sentence-Bert model (Reimers, 2019) to calculate the similarity between the generated research questions and standard answers under both conditions to evaluate the quality of the generated research questions. The comparison of generation results from different LLMs is shown in Table 6.

This paper sets the similarity threshold c = 0.8 as the accuracy threshold for generating research questions. Specifically, when c ≥ 0.8, the generated research question is correct; otherwise, the generated

Table 5: The results of the fine-tuning LLMs.

LLMs	BLEU	brevity_penalty	length_ratio	rouge1	rouge2	rougeL
Mistral-7B	0.167	0.939	0.941	0.506	0.255	0.449
Baichuan2-7B	0.121	0.980	0.980	0.459	0.215	0.405
Chatglm3-13B	0.133	0.982	0.982	0.463	0.213	0.411
Internlm-7B	0.161	0.948	0.950	0.493	0.243	0.429
Llama3-8B	0.154	0.932	0.934	0.491	0.239	0.430

Table 6: Comparison of generation results from different LLMs.

Instruction	Given the title, the research question and the research methods, can we distill a concise question summarizing the research issue addressed in this article? Please use appropriate question words! the title: 'DiffMatch: Diffusion Model for Dense Matching', the research question:'To address this, we propose DiffMatch, a novel conditional diffusion-based framework designed to explicitly model both the data and prior terms.', the research methods: 'Unlike previous approaches, this is accomplished by leveraging a conditional denoising diffusion model.'
Mistral	How can conditional diffusion models enable dense matching?
Baichuan2	How can diffusion models enable dense matching of images?
Chatglm3	How can conditional diffusion modeling enable dense matching?
Internlm	How effectively can conditional denoising diffusion models model both data and prior terms for dense matching?
Llama3	How accurately matches dense data a conditional denoising diffusion model?

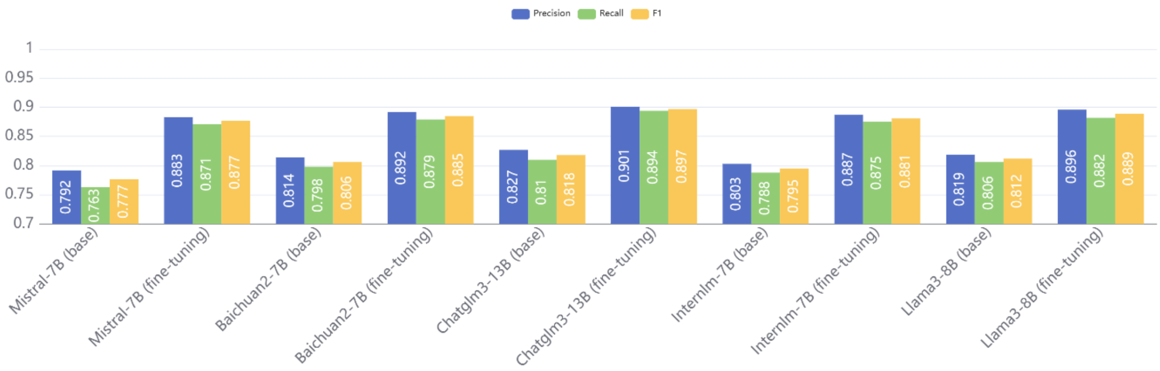

Figure 4: The experimental results of the benchmark LLMs and fine-tuned LLMs.

research question is incorrect. The experimental results of the benchmark models and fine-tuned models are shown in Figure 4. The experimental results demonstrate that the research question generation method based on meta-knowledge prompts provides more accurate and rich knowledge element inputs, reduces the difficulty of the recognition task, and improves the quality of research question generation.

5 CONCLUSIONS

This paper proposes a research question generation method based on meta-knowledge prompt engineering. To extract key meta knowledge required for generating research questions from scientific literature, a sentence classification model based on feature word vectors is proposed. Then, research question generation prompts that integrate meta-knowledge are used to fine-tune LLMs, which provide more accurate and targeted input, thereby improving the quality and accuracy of the generated results. The key contributions of this study are summarized as follows:

(1) In meta-knowledge extraction, we construct feature word sets for research objective sentences and research method sentences, and considers the feature word vector based on syntactic structure features. Utilizing the feature word vectors and the constructed. By concatenating the feature word vectors with the model's output, the model is trained, which helps model to capture and enhance the semantic expression and contextual information of feature words. Experimental results show that the DeBERTa model based on feature word vectors proposed in this paper achieves the best meta-knowledge extraction performance, with an F1 score of 0.97; compared to the original DeBERTa, the precision and recall are improved by 2.6% and 1.7%, respectively.

(2) Based on the key meta-knowledge: research objective sentences and research method sentences, research question prompts that integrate meta- knowledge are manually constructed, and LLMs are fine-tuned. Experimental results indicate that, the proposed method that integrates meta-knowledge extraction effectively improves the quality of generation, with an average F1 score of 88.6% after fine-tuning, an increase of 8.4%; from an individual model analysis, the fine-tuned Chatglm3-13B achieves the highest F1 score of 89.7%.

(3) This method can be applied to the generation task of research question sentences in different domains. In addition, by updating or replacing the meta-knowledge, it can generate different types of sentences, thereby providing a theoretical basis or model foundation for other downstream tasks.

Notably, this paper only optimizes the task of generating research question sentences for scientific literature. In future research, we plan to enhance the generation of other types of sentences. In addition, with the development of MultiModal LLMs, to improve the performance of text generation, combining multimodal data (such as images, tables, etc.) with prompt engineering is also one of the hot issues.

ACKNOWLEDGEMENTS

This study was funded by the National Key R&D Program of China(2022YFF0711900).

REFERENCES

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press. https://doi.org/10.7208/chicago/9780226458106.001.0001

Sarawagi, S. (2008). Information extraction. Foundations and Trends® in Databases, 1(3), 261-377.

Martinez-Rodriguez, J. L., Lopez-Arevalo, I., & Rios-Alvarado, A. B. (2018). Openie-based approach for knowledge graph construction from text. *Expert Systems with Applications*, *113*, 339-355.

Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.* https://doi.org/10.18653/v1/N19-1423

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.https://arxiv.org/abs/1907.11692

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Chen, Y., Zhang, Y., Hu, C., & Huang, Y. (2021). Jointly extracting explicit and implicit relational triples with reasoning pattern enhanced binary pointer network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5694-5703).

OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., ... & Tang, J. (2023). GLM-130B: An Open Bilingual Pre-trained Model. *arXiv preprint arXiv:2210.02414*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1-35.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1-7).

Alvesson, M., & Sandberg, J. (2013). Constructing research questions: Doing interesting research.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., ... & Barabási, A. L. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Safaya, A., Abdullatif, M., & Yuret, D. (2020). Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. *arXiv preprint arXiv:2007.13184*.

Li, X., Zhang, Z., Liu, Y., Cheng, J., Tian, X., Wang, S., Su, X., Wang, R., & Zhang, T. (2023). A study on the method of identifying research question sentences in scientific articles. *Library and Information Service*, 67(09), 132-140. https://doi.org/10.13266/j.issn.0252-3116.2023.09.014

Mei, X., Wu, X., Huang, Z., Wang, Q., & Wang, J. (2023). A multi-scale semantic collaborative patent text classification model based on RoBERTa. *Computer Engineering & Science*, 45(05), 903-910.

Su, J., Lu, Y., Pan, S., Wen, B., & Liu, Y. (2021). RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864.*

Devillers, M., Saulnier, P., Scialom, T., Martinet, J., Matussière, S., Parcollet, T., ... & Staiano, J. (2023). Mistral: A Strong, Efficient, and Controllable Multi-task Language Model. *arXiv preprint* arXiv:2304.08582. https://arxiv.org/abs/2304.08582

Wu, J., Li, D., Li, S., Fu, T., Chen, K., Wang, C., ... & Zhang, Z. (2023). Baichuan 2: Open Large-scale Language Models. *arXiv preprint* arXiv:2304.09070. https://arxiv.org/abs/2304.09070

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., ... & Tang, J. (2022). Glm-130b: An open bilingual pre-trained model.*arXiv preprint arXiv:2210.02414*.

Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., ... & Lin, D. (2024). Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Reimers, N. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.