Enhancing Cross-Lingual Dialogue Summarization through Interpretable Chain-of-Thought

Zhongtian Bao[1,2][0000−0001−6618−5325], Zhang Yao[1,2][0009−0003−3071−0513], Jun Wang[3][0000−0001−8932−6661], Adam Jatowt[4][0000−0001−7235−0665], and Zhenglu Yang [1,2][0000−0001−9528−965X]

[1] TMCC, TBI Center, CS, Nankai University, Tianjin, China
[2] Key Laboratory of Data and Intelligent System Security (DISSec), Ministry of Education, China
[3] Ludong University, Shandong, China
[4] University of Innsbruck, Austria
{1120200192, yaozhang, junwang}@mail.nankai.edu.cn, adam.jatowt@uibk.ac.at, yangzl@nankai.edu.cn

Abstract. The rapid development of large language model techniques in recent years has made effective summarization of cross-lingual dialogue information possible, which is crucial in today's global communication landscape. However, existing approaches often face problems with the lack of interpretability information and intermediate result analysis for the summarization generation process. In this paper, we propose two optimizations to address these issues. First, we use a self-reply analysis structure to extract the subtle attitude changes for each participant through the dialogue progress. Second, we combine this information to generate more interpretability cross-lingual dialogue summarization results. We propose a view-aware, chain-of-thought-based structure to clarify the generation process of cross-lingual dialogue summarization. The temporal properties of dialogue applications are considered throughout the computational process within our framework. Experimental results on cross-lingual summarization tasks in English, French, Spanish, Russian, Chinese, and Arabic, as well as cross-lingual hybrid tasks, demonstrate that our proposed method outperforms state-of-the-art baselines.

Keywords: Cross-lingual Dialogue Summarization · Chain-of-Thought · View-Aware Structure.

1 Introduction

Cross-lingual summarization is a critical task in natural language processing that aims to compress the information from a document context written in one language (e.g., English) and generate a coherent summary in another language (e.g., Chinese). Recently, the exploration of CLS has been extended to more complicated situations, such as dialogues where participants communicate using

their native languages. We propose to define this task as Cross-Lingual Dialogue Summarization (CLDS). The ability to effectively summarize cross-lingual dialogue information has become increasingly important in today's interconnected world, where communication often transcends linguistic boundaries. As information technology continues to advance rapidly, the significance of CLDS has been amplified, becoming a vital tool for facilitating understanding and collaboration in diverse contexts.

The demand for effective dialogue summarization has grown, particularly in multinational conferences, where speakers communicate in their native languages [14, 22]. In these settings, the specificity of feedback and the clarity of communication are essential. However, research that effectively integrates CLDS with dialogue summarization remains relatively unexplored. Many existing studies fail to adequately address the complexities that arise when different speakers use their mother languages to interact, leading to challenges in producing accurate and contextually relevant summaries. The specificity of CLDS is closely related to machine translation (MT) and mono-lingual summarization (MS). Thus, the CLDS domain has benefited from focused research efforts that refine methodologies and data sources tailored to cross-lingual contexts [3, 4, 18].

Recent advancements in large language models (LLMs), such as mBart [6] and mT5 [21], have shown promising potential in enhancing CLDS performance, leveraging vast amounts of training data to improve their understanding of linguistic nuances and contextual relationships. Despite the impressive performance of LLMs, researchers face significant challenges, particularly regarding the substantial computational resource [15]. Traditional approaches commonly involve a "pre-train, fine-tune, and predict" paradigm, which can be resource-intensive, time-consuming, and unexplainable. Researchers have explored effective and efficient methodologies to address these issues, transitioning to a "pre-train, prompt, and predict" paradigm. This innovative strategy allows for the reformulation of downstream tasks to resemble those encountered during the original training of the language model, utilizing carefully designed prompts to guide the model's output. This flexibility enables researchers to manipulate model outputs effectively, facilitating desired predictions even in zero-shot scenarios [7, 9, 15], and leading to superior performance in cross-lingual summarization in recent years [5, 10].

While advancements have been made, two persistent issues remain in CLDS: the challenge of interpretability and the need for explanations of summarization generation, as illustrated in Fig. 1. Participants' viewpoints are changing throughout continuous communication, and emphasizing this temporal aspect of dialogue summarization. Furthermore, existing state-of-the-art approaches often overlook the potential benefits of interpretability information in the summarization process, which are essential for clarifying source information and demonstrating the soundness of the summaries.

To effectively capture and utilize complex or intricate information, we introduce a self-guidance structure to ensure interpretability. With the assistance of GPT, we improve the tracking of information flow in cross-lingual dialogue

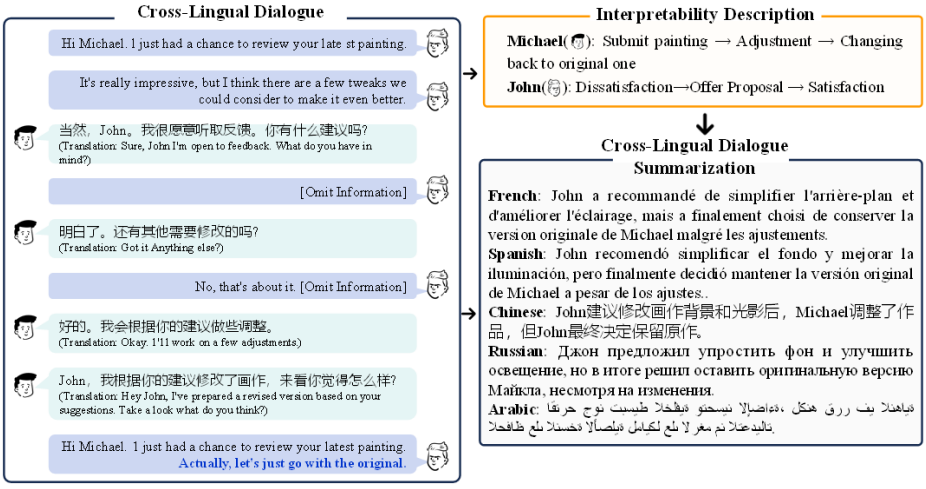

Fig. 1. The task overview of CLDS. Different speakers use their mother languages to communicate with each other. During this process, the interpretablity information plays an important part in generating appropriate summarization.

scenarios. We also propose a novel view-aware, chain-of-thought-based prompt model. As demonstrated in our framework, this approach facilitates self-guidance by integrating information from various perspectives throughout the dialogue summarization process. Leveraging a chain-of-thought framework enhances the interpretability and effectiveness of cross-lingual dialogue summarization.

Our main contributions are as follows:

(1) We propose an innovative view-aware, chain-of-thought-based structure for the CLDS task. This structure accounts for the differing specificity aspects within cross-lingual hybrid dialogue scenarios. Enhancing interpretability enables the chain-of-thought framework to understand the inference process better.

(2) A preliminary empirical study demonstrates the superior performance of our model across various evaluations in the CLDS task. Furthermore, our model outperforms the competing approaches in interpretability.

2 Related Work

CLDS consists of two sub-tasks: cross-lingual summarization (MS) and cross-lingual translation (MT). Traditional CLDS methods are pipeline-based, treating MS and MT as independent processes. These methods can be broadly categorized into two approaches: first summarization followed by translation (Sum-Trans) and first translation followed by summarization (Trans-Sum) [13, 19]. While these methods have clear interpretability, the pipeline strategy often struggles to achieve optimal performance due to error accumulation between the two steps.

To bridge the gap between the translation and summarization steps, researchers are increasingly focusing on End-To-End (End2End) methods. These approaches fine-tune LLMs to directly generate target language summaries using the Sequence-to-Sequence (Seq2Seq) architecture, which leads to significant improvements in cross-lingual summarization tasks. In addition, other strategies have been developed to incorporate with End2End for enhanced performance, such as multi-task learning, specific word attention, knowledge distillation, and variational auto-encoders [2, 12].

Despite the high achievements obtained by fine-tuning LLMs, neural network parameters are still optimized by gradient descent, which causes high computational consumption and needs sufficient data sources. To alleviate this issue, the prefix tuning strategy that only adjusts a part of the parameters in LLMs is often utilized [1]. Based on this strategy, prompt-based methods that use prompts to guide LLMs can reduce time complexity and generate high-quality target summaries. The main idea of prompt-based methods is to design an input format to help LLMs use the general language information learned through the pre-training stage to solve downstream tasks. The prompt and prompt-based chain-of-thought approaches have shown their potential in the cross-lingual summarization domain with the assistance of LLMs [16, 20]. Moreover, these methods have shown promising research space in zero-shot and few-shot situations.

Providing a clear explanation of the summarization process is often necessary. In this paper, we introduce a noise-tolerant self-reply module and a temporal information analysis module designed to improve interpretability.

3 Our Framework

Our proposed framework consists of three components, as portrayed in Fig. 2. In the central part of the framework, we employ a chain-of-thought prompt structure to decompose the CLDS task into summarization and translation subtasks. On the right side, we propose an interpretability module to calculate the temporal weight of different sentences, capturing shifts in viewpoints among various participants during the communication. In summary, our framework aims to enhance CLDS through cross-lingual-focus-aware chain-of-thought structure, and interpretability modeling.

3.1 Preliminaries

Cross-Lingual Dialogue Summarization Task The CLDS task is to read a source document in language A, denoted as $D^A = \{x_1^A, x_2^A, ..., x_m^A\}$, to generate a summary in language B, represented as $S^B = \{y_1^B, y_2^B, ..., y_n^B\}$, where m is the word token number of input dialogue and n is the length of summary. The target language summary S^B can be generated through the following optimization function:

$$\arg\max_{S^B} P(S^B|D^A). \tag{1}$$

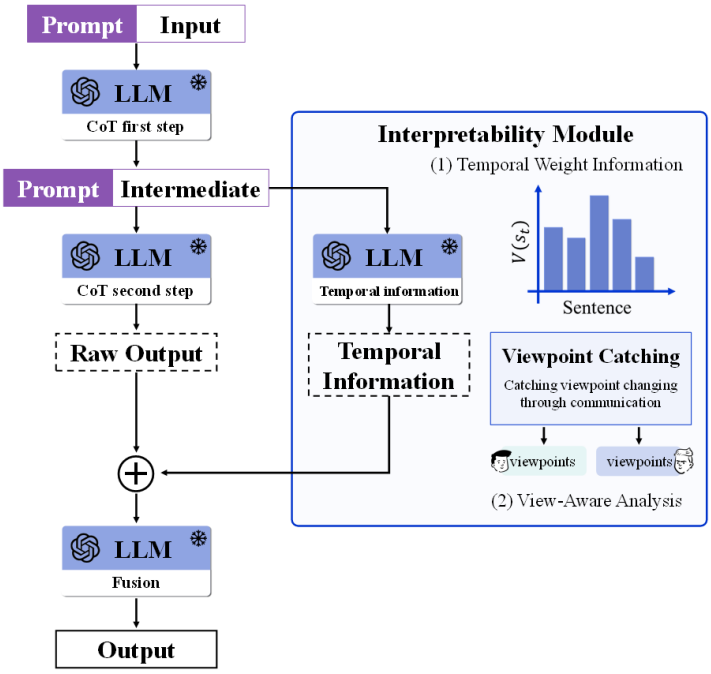

Fig. 2. Overview of the proposed structure. The purple parts symbolize that their parameters need to be trained, while the white parts symbolize that their parameters are frozen. Initially, we separated the CLDS task into different paths to catch different focus information. In Part (1), we analyze the temporal information, and in Part (2) to examine the view-related changes through communication.

The solution of Eq. (1) is disparate across different models. For example, in End2End-based Seq2Seq CLDS models, the conditional probability can be induced through the assistance of *Encoder* and *Decoder* as follows:

$$\begin{aligned} Loss &= -\sum_{i=1}^{n} \log P(y_i^B | h, S_{<i}^B), \\ P(y_i^B | h, S_{<i}^B) &= Decoder(h, S_{<i}^B), h = Encoder(D^A). \end{aligned} \tag{2}$$

Prompt Structure Fine-tuning methods use gradient descent algorithms to adjust network parameters while optimizing Eq. (2). This back-propagate learning method allows language models trained by upstream tasks to be fine-tuned to downstream tasks with high time consumption features and sizeable computational complexity. To overcome this disadvantage, we use fixed hard prompting that directly inserts human-designed sentences at the beginning of the cross-lingual dialogue sentences to instruct the detailed task information.

Specifically, we insert the instruct sentence or trainable vector p at the beginning of dialogue sentence D^A when constructing the prompt-tuning structure. The summary S^B can be generated as follows:

$$P(S^B|D^A) = \prod_{i=1}^{n} P(S_i^B|D'^A, S_{<i}^B), D'^A = [p; D^A]. \tag{3}$$

3.2 Chain-of-Thought Prompt Structure

Despite the prompt method's convenience, its performance in complicated tasks remains underexplored, such as CLDS, which is rarely considered in the pretrained progress. For example, one-step prompting methods easily fall into the hallucination that directly generates translation results for input dialogue sentences without summarization progress, especially in Chinese, Russian, and Arabic-related CLDS situations.

To overcome this deficit, we raise a chain-of-though prompt structure to disassemble the whole task into trans-sum and sum-tran steps, similar to the pipeline methods. With the assistance of contextual information, our structure combines the intermediate results, which can give LLM more convincing and detailed instruction prompting and reduce the error accumulated between translation and summarization steps.

For the tran-sum path, we use the GPT API to translate the input dialogue to the target language B, represented as $\tilde{D}^B$, which m is the length of intermediate result $\tilde{D}^B$. The final CLDS result can be generated through the following formulation:

$$P(\tilde{D}^B|X_{Tran}, D^A) = \prod_{i=1}^{m} P(\tilde{D}_i^B|X_{Tran}, D^A, \tilde{D}_{<i}^B), \tag{4}$$

$$\log P(S^B|X_{Sum}, \tilde{D}^B, D^A) = \log \sum_{i=1}^{n} P(S_i^B|X_{Sum}, \tilde{D}^B, D^A, S_{<i}^B). \tag{5}$$

For the sum-tran path, we use GPT-related API to summarize the input dialogue, represented as $\tilde{S}^A$, which m is the length of intermediate result $\tilde{S}^A$. The final CLDS result can be generated through the following formulation:

$$P(\tilde{S}^A|X_{Sum}, D^A) = \prod_{i=1}^{m} P(\tilde{S}_i^A|X_{Sum}, D^A, \tilde{S}_{<i}^A), \tag{6}$$

$$\log P(S^B|X_{Tran}^{En}, \tilde{S}^A, D^A) = \log \sum_{i=1}^{n} P(S_i^B|X_{Sum}, \tilde{S}^A, D^A, S_{<i}^B). \tag{7}$$

3.3 Interpretability Module

A major disadvantage troubling researchers is the black-box character of LLMs. In the chain-of-thought structure, we separate CLDS into summarization and

translation sub-tasks to better understand what happens in the intermediate steps. We propose the temporal weight and view-aware analysis modules to better catch the interpretability information and generate related details.

Temporal Weight In contrast to document summarization, which may use the literature skill of inverted order or narration interspersed with flashbacks, dialogue summarization generally focuses on participants' communication. During the communication, different participants' viewpoints may change dynamically. We follow the Temporal-Difference (TD) algorithm to capture this temporal information [11, 17]. TD uses current state values through future states in the reinforcement learning analysis.

Given the sentence value in time t as s_t, we use the LLM to evaluate s_t, represented as $V()$. We utilize the following function to update the TD value $V(s_t)$:

$$V(s_t) \leftarrow V(s_t) + \alpha(r_{t-1} + V(s_{t-1}) - V(s_t)), \tag{8}$$

where α is the learning rate, and r_{t-1} is the weight of information gain in time $t-1$, which is generated by the GPT API.

View-Aware Analysis Participants' attitudes and views change during the communication; hence, we propose the view-aware prompting structure that uses chain-of-thought as the extra information to give the GPT API better self guidance. During this progress, we limit the viewpoint to several words, which can guide GPT to understand the skeleton of the whole communication progress. Then, we use the task description T_{view} to extract the information $\tilde{S}_{View}$ as

$$\mathrm{GPT}(S, T_{View}) \rightarrow \tilde{S}_{View}. \tag{9}$$

4 Experiments

4.1 Dataset

We conduct our experiments on the MSAMSum [8]. To better demonstrate the generalization and robustness of our proposed model, we conducted corresponding experiments on the two aforementioned datasets.

MSAMSum comprises six languages: English, French, Spanish, Chinese, Russian, and Arabic. The MSAMSum dataset uses a round-trip translation strategy with quality control to gather the cross-lingual dialogue summarization with textual alignment. MSAMSum has five different sub-tasks to quantify the effectiveness of CLDS, which are One-To-One, Many-To-One, One-To-Many, and Mix-To-Many tasks. The language-supported situation classifies these tasks during input and output progress. We use the Mix-To-Many task to test the CLDS performance, which hybrids different languages with different speakers speaking their languages.

Table 1. Zero-Shot CLDS performance under the F1 score of ROUGE in the Mix-To-Many task under the English, French, Spanish, Russian, Chinese, and Arabic parts of the MSAMSum dataset: ROUGE-1, ROUGE-2, and ROUGE-L are represented by R-1, R-2, and R-L, respectively.

Model	**English**			**French**			**Spanish**		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mT5 fine-tuning	.2587	.0674	.1843	.2453	.0598	.1820	.2395	.0571	.1836
GPT-3.5 End2End	.3266	.1005	.2474	.3197	.1213	.2468	.3283	.1163	.2485
GPT-4o End2End	.3284	.1110	.2492	.3245	.1304	.2455	.3285	.1291	.2519
GPT-4o Tran-Sum	.3321	.1101	.2537	.3392	.1415	.2602	.3354	.1298	.2564
MoP	.3354	.1143	.2541	.3350	.1417	.2595	.3348	.1276	.2553
MoP w Interpretability	**.3421**	**.1240**	**.2610**	**.3420**	**.1423**	**.2615**	**.3395**	**.1305**	**.2640**
Model	**Arabic**			**Russian**			**Chinese**		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
mT5 fine-tuning	.1226	.0147	.0732	.2057	.0432	.1576	.2567	.1042	.2031
GPT-3.5 End2End	.1753	.0381	.1491	.2777	.0715	.2284	.3150	.1515	.2466
GPT-4o End2End	.1820	.0346	.1584	.2767	.0785	.2226	.3303	.1673	.2573
GPT-4o Tran-Sum	.1917	.0406	.1648	.2882	.0815	.2320	.3542	.1854	.2785
MoP	.1905	.0358	.1594	.2876	.0809	.2375	.3507	.1810	.2742
MoP w Interpretability	**.1945**	**.0421**	**.1698**	**.2919**	**.0805**	**.2378**	**.3601**	**.1894**	**.2802**

4.2 Experimental Settings

We evaluate our experiments using the GPT-3.5-turbo and GPT-4o API configuration. In addition to GPT-based methods, we use the mT5 model as the baseline. We fine-tune the mT5 model on the MSAMSum dataset using an end-to-end approach for downstream tasks. This setup directly maps input sequences to the target outputs;

We utilize the Adam Optimizer to adjust the entire model structure with parameters $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-9}$. We implement a beam search with a beam size of four and a repetition penalty of 5.0 for evaluations. We used the hyperparameter of temperature as 0.75 and set the top-p value in GPT API as 0.9. These experiments provide comparative insights into the effectiveness of state-of-the-art pre-trained language models when applied directly to the task.

4.3 Experimental Results

We evaluate the performance of all models using the standard ROUGE metric and report F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. Table 1 displays the results of mix-to-many sub-tasks in the MSAMSum dataset. The data shows that our proposed structure, which using interpretability information, significantly surpasses the baseline mT5 and chain-of-thought structure. Moreover, our proposed chain-of-thought structure is well-suited for all six languages: English, French, Spanish, Arabic, Russian, and Chinese. In the experiments conducted in this study, we improved the ROUGE-1 score from 0.2587 to 0.3421 by interpretability information. Our analysis indicates that the interpretability

components complemented each other during the experiments, contributing to the results.

5 Conclusion

This paper introduces a novel CLDS framework that utilizes a chain-of-thought prompt structure designed to enhance interpretability information. By implementing this framework, we can effectively capture temporal information and shifts in perspective throughout the dialogue, which is particularly crucial in complex scenarios. The comparative experiments conducted in this study clearly demonstrate that our model outperforms baseline models in various performance metrics, highlighting the advantages of our approach. Additionally, we believe that by focusing on the intricate dynamics of conversation, we can further refine the system's ability to understand context and user intent. In future work, we intend to delve deeper into the concept of view-aware information, as we believe this will significantly contribute to improving both the performance and interpretability of our framework.

Acknowledgements: This work was supported in part by the National Natural Science Foundations of China (Nos. 62306156, 62106091), Fundamental Research Funds for the Central Universities, Nankai University (No. 63241436).

References

1. Bhatnagar, N., Urlana, A., Mujadia, V., Mishra, P., Sharma, D.M.: Automatic data retrieval for cross lingual summarization. In: ICNLP. pp. 822–827 (2023)
2. Cai, Y., Yuan, Y.: CAR-Transformer: Cross-attention reinforcement transformer for cross-lingual summarization. In: AAAI. pp. 17718–17726 (2024)
3. Cao, Y., Liu, H., Wan, X.: Jointly learning to align and summarize for neural cross-lingual summarization. In: ACL. pp. 6220–6231 (2020)
4. Cao, Y., Wan, X., Yao, J., Yu, D.: MultiSumm: towards a unified model for multilingual abstractive summarization. In: AAAI. pp. 11–18 (2020)
5. Chen, Y., Zhang, H., Zhou, Y., Bai, X., Wang, Y., Zhong, M., Yan, J., Li, Y., Li, J., Zhu, M.: Revisiting cross-lingual summarization: A corpus-based study and a new benchmark with improved annotation. arXiv preprint arXiv:2307.04018 (2023)
6. Chipman, H.A., George, E.I., McCulloch, R.E., Shively, T.S.: mBART: Multidimensional monotone BART. Bayesian Anal. **17**(2), 515–544 (2022)
7. Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H.T., Sun, M.: Openprompt: An open-source framework for prompt-learning. In: ACL. pp. 105–113 (2022)
8. Feng, X., Feng, X., Qin, B.: MSAMSum: Towards benchmarking multi-lingual dialogue summarization. In: ACL. pp. 1–12 (2022)
9. Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L.: Pre-trained models: Past, present and future. AI Open **2**, 225–250 (2021)
10. Le, T.: Cross-lingual summarization with pseudo-label regularization. In: NAACL. pp. 4644–4677 (2024)

11. Li, X., Deng, Z.D., Rauchenstein, L.T., Carlson, T.J.: Contributed review: Source-localization algorithms and applications using time of arrival and time difference of arrival measurements. Rev. Sci. Instrum. **87**(4) (2016)
12. Liang, Y., Meng, F., Zhou, C., Xu, J., Chen, Y., Su, J., Zhou, J.: A variational hierarchical model for neural cross-lingual summarization. arXiv preprint arXiv:2203.03820 (2022)
13. Lin, Y., Liu, Z., Sun, M.: Neural relation extraction with multi-lingual attention. In: ACL. pp. 34–43 (2017)
14. Liu, N., Wei, K., Yang, Y., Tao, J., Sun, X., Yao, F., Yu, H., Jin, L., Lv, Z., Fan, C.: Multimodal cross-lingual summarization for videos: A revisit in knowledge distillation induced triple-stage training method. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10697–10714 (2024)
15. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
16. Ma, J., Huang, Y., Wang, L., Huang, X., Peng, H., Yu, Z., Yu, P.: Augmenting low-resource cross-lingual summarization with progression-grounded training and prompting. ACM T. Asian and Low-Reso. **23**(9), 1–22 (2024)
17. Myers, E.W.: An o (nd) difference algorithm and its variations. Algorithmica **1**(1), 251–266 (1986)
18. Wan, X., Luo, F., Sun, X., Huang, S., Yao, J.g.: Cross-language document summarization via extraction and ranking of multiple summaries. Knowl. Inf. Syst. **58**, 481–499 (2019)
19. Wang, J.D., Chang, D., Meng, F.Q., Qu, G.: A comprehensive survey and prospect of cross-lingual summarization method research. J. Netw. Intel. **9**(1), 384–412 (2024)
20. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS. pp. 24824–24837 (2022)
21. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A massively multilingual pre-trained text-to-text transformer. In: NAACL-HLT. pp. 483–498 (2020)
22. Zhu, J., Wang, Q., Wang, Y., Zhou, Y., Zhang, J., Wang, S., Zong, C.: NCLS: Neural cross-lingual summarization. In: EMNLP-IJCNLP. pp. 3054–3064 (2019)