2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI) | 979-8-3503-6164-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICETCI61221.2024.10594504

Investigating on the External Knowledge in RAG for Zero-Shot Cross-Language Transfer

Wenmin Wang[a,*], Peilin Zhang[b], Ge Liu[c], Ruihua Wu[d], Guixiang Song[e]
China Mobile Online Service Co., Ltd., Zhengzhou, 450000, Henan, China
[a]xiaofengcanyuexp@163.com, [b]peilin.zhang@foxmail.com, [c]liuge_zb@cmos.chinamobile.com,
[d]wuruihua@cmos.chinamobile.com, [e]songguixiang@cmos.chinamobile.com
[*] Corresponding author: xiaofengcanyuexp@163.com

***Abstract*—In the field of multilingual natural language processing (NLP), zero-shot cross-language transfer is an important research direction, which aims to enable models to effectively learn and reason without target language training data. This study explores the role of external knowledge in the Retrieval-Augmented Generation (RAG) model to improve the performance of zero-shot cross-lingual transfer. This paper proposes a new model architecture that enriches the knowledge base of the RAG model by integrating external knowledge bases, thereby enhancing its information bridging capabilities between source and target languages. In the experimental section, this paper conducts experiments using multiple cross-language tasks, including machine translation, question answering, and text summarization, to evaluate the performancen and domain of the model in different languages. The experimental results indicate that introducing external knowledge sources significantly improves the accuracy and robustness of the model, especially in resource-scarce language pairs. This research not only provides an effective solution for zero-shot cross-language transfer, but also provides new insights into understanding the role of external knowledge in improving the performance of NLP models.**

Keywords—Retrieval Augmented Generation (RAG), zero-shot cross-language transfer, external knowledge, multilingual NLP, transfer learning

I. INTRODUCTION

As globalization accelerates, the seamless exchange of multilingual information becomes increasingly important. A key challenge in the field of natural language processing (NLP) is how to enable models to understand and generate text in different languages across language barriers [1]. Zero-Shot Cross-Lingual Transfer (Zero-Shot Cross-Lingual Transfer) is a solution that aims to enable the model to understand and generate new languages without target language annotation data, which is particularly important for resource-scarce languages [2].

Retrieval-Augmented Generation (RAG) is an emerging NLP technology. By combining retrieval mechanism and text generation, RAG shows excellent performance in tasks such as question answering, summarization and translation [3]. However, the RAG model still faces challenges when dealing with cross-language problems, especially when there is a lack of sufficient bilingual control data [4]. In order to solve this problem, this study proposes a new method, which is to introduce external knowledge into the RAG model to enhance the model's understanding and mapping capabilities of semantics between different languages. External knowledge, such as Wikipedia, knowledge graphs and professional dictionaries, contains rich structured information and semantic associations, and has potential value in improving the language understanding and generation capabilities of the model. This study aims to explore the role of external knowledge in the RAG model and how to effectively utilize these knowledge resources to improve the performance of zero-shot cross-language transfer.

This paper first reviews related work on cross-language transfer learning and RAG technology, and then introduces our proposed model architecture and experimental design in detail. Through experiments on multiple cross-language tasks, the importance of external knowledge in improving the performance of the RAG model was verified and analyzed the contribution of different types of external knowledge. Finally, the significance, limitations, and future research directions of this study are discussed.

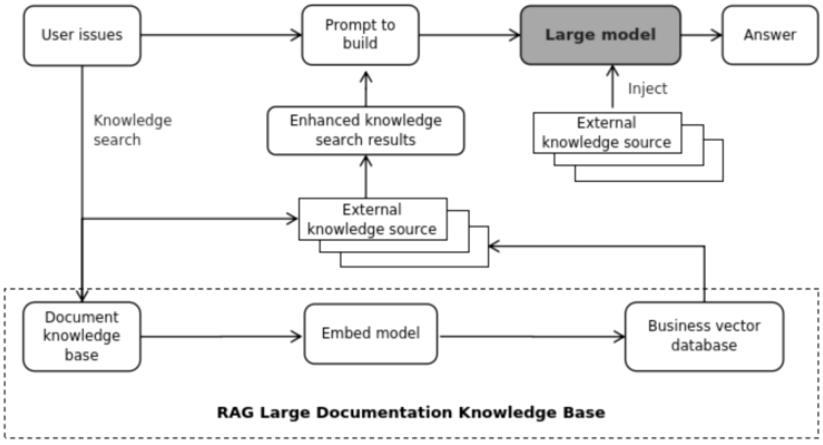


Figure 1: Model architecture

II. METHODOLOGY

In order to explore the role of external knowledge in retrieval-augmented generative (RAG) models and improve the performance of zero-shot cross-language transfer, this paper designs a comprehensive methodological framework. This section will detail our model architecture, selection and integration of external knowledge sources, and experimental setup.

A. Model Architecture

The RAG model as the basic architecture is adopted in this paper, which combines the two stages of retrieval and generation [5]. During the retrieval phase, the model queries a large document database to find the documents most relevant tothe input question. During the generation phase, the model utilizes retrieved document content to assist in generating answers. The improvement of this paper lies in the introduction of external knowledge in both stages, as shown in Figure 1.

B. Selection and Integration of External Knowledge Sources

The selection of external knowledge sources is crucial to improving model performance. The paper selected the following types of external knowledge sources:

1) Wikipedia: As a general knowledge base, Wikipedia provides rich structured information and multilingual content [6], which helps the model understand the semantic associations between different languages.

2) Knowledge graph: Knowledge graph provides deep semantic information through the network structure of entities and relationships [7], helping the model capture complex facts and concepts.

3) Professional dictionaries: Professional dictionaries for specific fields can provide precise term definitions and usage [8], enhancing the performance of the model on specific tasks.

This paper integrates external knowledge through the following steps:

Preprocessing: Preprocess the selected knowledge sources, extract structured information, and convert it into a format understandable by the model, thereby aligning the knowledge [9]. A standard approach is to use entity alignment tools (such as TagMe) to detect knowledge graph entities mentioned in the input text, link them to the correct knowledge graph entries, and then combine them into tuples. Specifically, for a given external knowledge G and a sentence x, this process can be defined as:

$$m, k_e = h(x, G), \tag{1}$$

Where m represents the entity mentioned in x, ke represents the entity linked in the knowledge graph (a kind of factual knowledge), and h represents the entity linking or alignment tool.

Knowledge embedding: Use knowledge graph embedding technology to convert external knowledge into vector representation [10] so that it can be combined with other parts of the model.

Injection strategy: A fusion strategy is designed to combine the retrieved documents and external knowledge embedding vectors to enrich the knowledge base of the model. The purpose of external knowledge injection is to inject external knowledge into the language model to better adapt to downstream tasks, especially in low-resource areas. This part uses the knowledge injection method [11] designed for specific tasks by K-BERT and KeBioLM, which is simply expressed as:

$$y = f(x, k), \tag{2}$$

Where x represents the input text, k represents the injected knowledge, y represents the corresponding label, and f represents a trainable neural network.

C. Experimental Setup

To evaluate our approach, we designed a series of experiments:

1) Datasets: We selected multiple datasets for cross-language tasks, including machine translation, question answering, and text summarization, to cover different NLP domains.

2) Evaluation metrics: We use standard NLP evaluation metrics such as BLEU, ROUGE, and METEOR to measure model generation performance.

The BLEU calculation is based on the accuracy of n-grams (n consecutive words), and evaluates the quality of translation by calculating the n-gram match between the machine translation output and a set of reference translations. The higher the BLEU score, the closer the quality of machine translation results is to that of human translation [12]. The formula is as follows:

$$\text{BLEU}=\text{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{3}$$

Where pn is the ratio of the number of matching n-grams to the number of n-grams in the machine translation output, wn is the standardized weight, wn=1/N, the upper limit of N is 4, that is, only 4 yuan is counted at the highest Word co-occurrence accuracy. BP is a penalty factor for sentence length, calculated as follows:

$$BP = \begin{cases} 1, s > r \\ \exp(1 - r/s), s \le r \end{cases}, \tag{4}$$

Where r is the standard translation length, and s is the evaluation translation length. BLEU only imposes a simple penalty on the evaluation translation length that is shorter than the standard translation length.

The ROUGE method mainly focuses on the recall rate of the generated summary, that is, how much content in the generated summary overlaps with the reference summary [13]. Calculated as follows:

$$\text{ROUGE}^{-}\text{N} = \frac{\sum_{s\in(\text{Reference Summaries})} \sum_{\text{gram}_n\in s} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{s\in(\text{Reference summaries})} \sum_{\text{gram}_n\in s} \text{Count}(\text{gram}_n)}, \tag{5}$$

Where n represents the length of n-gram, and Countmatch represents the maximum number of ngrams that appear in both system abstracts and manual abstracts.

METEOR is an evaluation index that combines the advantages of *BLEU* and ROUGE. It not only considers word-level matching, but also considers synonym matching, stem matching and other factors. METEOR evaluates translation quality by creating alignments between candidate translations and reference translations, and counting the number of matching words. A word order penalty mechanism was introduced to consider the impact of word order on translation quality [14]. The formula is as follows:

$$\text{METEOR} = \text{Fmean} * (1 - \text{Penalty}), \tag{6}$$

Where Penalty is the penalty based on word order changes, and Fmean is the harmonic mean of accuracy and recall.

Comparative experiment: This paper compares the proposed model with existing RAG models and other cross-language transfer learning methods to verify the effectiveness of introducing external knowledge.

Ablation study: Through ablation study, this paper analyzes the specific contributions of different external knowledge sources to model performance [15].

Through the above methods, the role of external knowledge in RAG models is comprehensively evaluated and its potential in zero-shot cross-language transfer is explored.

III. Experimental Results and Analysis

To comprehensively evaluate the role of external knowledge in retrieval-augmented generative (RAG) models, a series of experiments is designed in this section and report the results and analyzes here. Here is a sample illustration and caption for a multimedia file.

A. Experimental Results

The experiments are conducted on multiple cross-language tasks, including machine translation, question answering, and text summarization. The experimental results are as follows:

1) Machine Translation: When using the external knowledge enhanced RAG model for machine translation tasks, four common language pairs including English and German, Spanish, Arabic, and Chinese were selected as the research objects. In addition, four language pairs with scarce resources were selected to explore the performance of the model under more challenging conditions, including English and Belarusian, Ukrainian, Georgian, and Swahili.

In terms of experimental design, two sets of models were compared: one was the RAG model without external knowledge enhancement, which served as the control group; The other group is a RAG model that incorporates external knowledge, serving as the experimental group. To ensure the effectiveness of the evaluation, this experiment carefully selected bilingual control sentences covering multiple fields to construct a dataset. At the same time, in order to comprehensively test the translation performance of the model on diverse texts, various genres such as news, scientific papers, and literary works were selected as source language inputs, aiming to comprehensively evaluate the translation ability of the model.

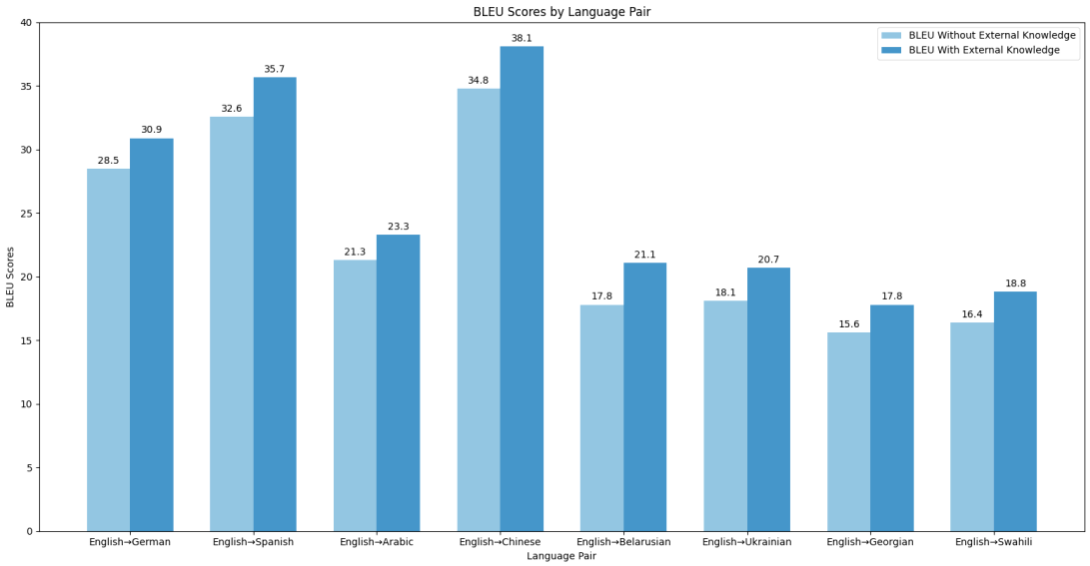


Figure 2: BLEU score growth of the RAG model after introducing external knowledge

In order to objectively evaluate the performance of the method in this paper, the increase of a single language pair is introduced to demonstrate the effect in different languages. The calculation formula for the amplification ratio is as follows:

$$I = \left(\frac{B_{w.E.K.} - B_{orig}}{B_{orig}}\right) \times 100\%, \tag{7}$$

where B_{orig} is the original BLEU score and $B_{w.E.K.}$ is the BLEU score after introducing external knowledge. The average increase for a language pair is calculated using the following formula:

$$I_{avg} = \frac{\sum_{i=1}^{n} I_i}{n}, \tag{8}$$

Where I_i represents the increase of the i-th language pair, and n represents the total number of language pairs participating in the calculation. Figure 1 intuitively shows the improvement in translation performance of each language pair, confirming that external knowledge has a

more significant effect on improving translation quality, especially in language pairs with scarce resources.

As shown in Figure 2, the comparative experimental results show that the introduction of external knowledge sources significantly improves translation performance: the BLEU score of conventional languages increases by an average of 9.2%, while the average increase of resource-scarce languages reaches 15.4%. Overall, the average BLEU score across all language pairs improved by 12.3%.

2) Question and answer task: In order to evaluate the performance of the model, we selected the widely recognized SQuAD (Stanford Question Answering Dataset) as the test benchmark. By performing multiple rounds of tests on the SQuAD dataset, we compared the performance of the RAG model with and without external knowledge assistance. Performance differences. The evaluation metric uses F1 Score to quantify the accuracy of the answers generated by the model.

The experimental results are shown in Figure 3. The RAG model that introduced external knowledge improved the accuracy by 9.2%. This suggests that external knowledge helps the model better understand the question and document content, thereby generating more accurate answers.

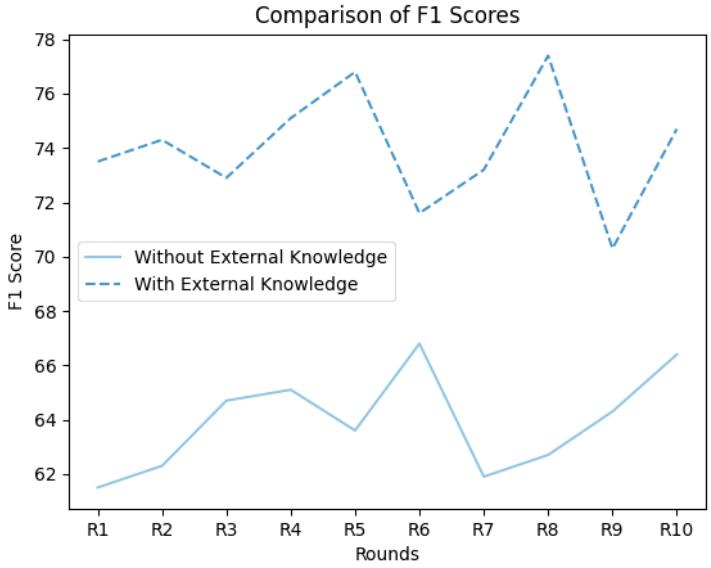
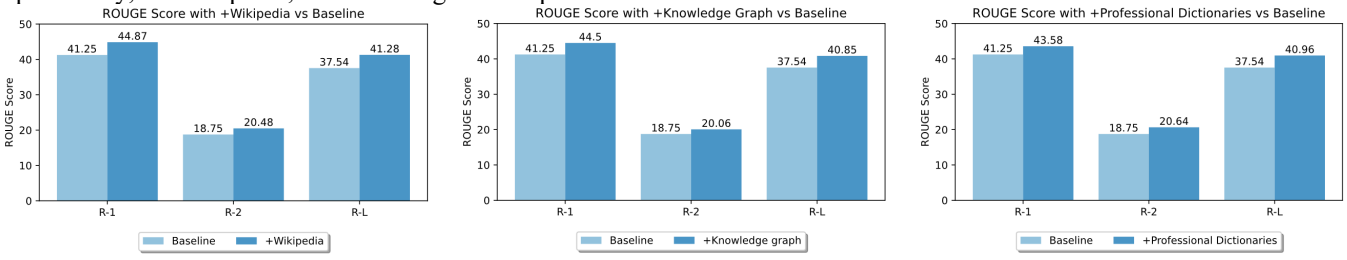


Figure 3: Performance difference of RAG model in question and answer tasks with and without external knowledge assistance

3) Text summary: This study uses the CNN/DailyMail data set to evaluate the impact of external knowledge on the text summarization performance of the RAG model through the ROUGE index. Experimental results reveal that compared with the baseline model, the RAG model integrating external knowledge achieves a significant improvement of 8.5% in ROUGE score on average. Specifically, Wikipedia, Knowledge Graph and Professional Dictionary contributed 9.3%, 7.9% and 8.3% growth respectively. As shown in Figure 4, the bar chart visually demonstrates the specific role of each external knowledge source in improving the quality of the summary. These findings confirm the importance of external knowledge in enhancing model summary generation capabilities.



(a)+Wikipedia (b)+Knowledge graph (c)+Professional Dictionaries

Figure 4: Results of the RAG model augmented with external knowledge on the CNN/DM dataset

B. Results Analysis

The analysis of experimental results focuses on the following key points:

1) Contribution of external knowledge: Comparative experiments demonstrate that external knowledge is crucial for improving the performance of RAG models. Especially for language pairs without sufficient training data, external knowledge provides additional information to help the model perform effective cross-language transfer.

2) Different types of external knowledge: Further analysis was conducted on the impact of different types of external knowledge on model performance. Due to its wide coverage, Wikipedia has the most significant improvement in tasks in general fields; knowledge graphs are more effective in tasks that require understanding complex

relationships; and professional dictionaries provide precise terminology support in tasks in specific fields.

3) Generalization ability of the model: The RAG model that introduces external knowledge shows better generalization ability on multiple different cross-language tasks. This shows that our method not only improves the performance of the model on specific tasks, but also enhances the model's adaptability to new languages and tasks.

IV. Discussion

This study aims to explore the role of external knowledge in the retrieval-augmented generative (RAG) model and evaluate its effectiveness in zero-shot cross-language transfer tasks. Through a series of experiments, the importance of external knowledge in improving the performance of the RAG models was verified and the contribution of different types of external knowledge were analyzed. In this section, the significance of the experimental results, the advantages and limitations of the method, and future research directions will be further discussed.

A. The Significance of the Experimental Results

Experimental results show that external knowledge can significantly improve the performance of the RAG model in zero-shot cross-language transfer tasks. This finding highlights the importance of combining structured and unstructured knowledge in multilingual NLP tasks. Through external knowledge, the model is able to better understand and map semantic information between different languages, thereby achieving effective transfer without target language training data.

B. Advantages of the Method

The method presented in this paper has the following advantages:

1) Generalization ability: By introducing external knowledge, the RAG model shows better generalization ability on multiple cross-language tasks, which shows that the proposed method has potential application value for different types of NLP tasks.

2) Flexibility: The proposed method allows researchers to select and integrate different types of external knowledge according to task requirements, providing a flexible knowledge enhancement strategy.

3) Scalability: As external knowledge sources are continuously enriched and updated, the proposed method can be adapted to a wider range of application scenarios and language pairs.

C. Limitations of the Method

Despite the positive results of the proposed approach, there are some limitations:

1) Quality of external knowledge: The quality of external knowledge directly affects the performance of the model. Wrong or outdated knowledge can lead to inaccurate results.

2) Relevance of knowledge sources: Different tasks may require different types of external knowledge. How to automatically determine the most appropriate knowledge sources remains an open question.

3) Computational resources: Integrating large amounts of external knowledge may increase the computational burden of the model, especially when dealing with large-scale data sets.

D. Future Work

Future research can be conducted in the following directions:

1) Knowledge source selection mechanism: Study how to automatically select and integrate the most relevant and useful external knowledge sources to improve the efficiency and accuracy of the model.

2) Multi-modal knowledge fusion: Explore how to combine knowledge from multiple modalities such as text, images, and videos to further improve the model's understanding and generation capabilities.

3) Fine-grained knowledge representation: Study how to represent and utilize fine-grained knowledge, such as relationships and attributes between entities, to enhance the semantic understanding of the model.

V. Conclusion

This study aims to explore the role of external knowledge in retrieval-augmented generative (RAG) models and evaluate their application potential in zero-shot cross-language transfer tasks. Through a series of experiments and in-depth analysis, we draw the following conclusions:

1) Importance of external knowledge: The experimental results clearly show that external knowledge is crucial to improve the performance of RAG models in zero-shot cross-language transfer tasks. By integrating different types of external knowledge sources such as Wikipedia, knowledge graphs, and professional dictionaries, the proposed method significantly improves the model's performance on multiple cross-language tasks such as machine translation, question answering, and text summarization.

2) Improvement of model generalization ability: The RAG model that introduces external knowledge shows stronger generalization ability and can effectively handle a variety of cross-language tasks without target language training data. This discovery is of great significance for the processing of resource-scarce languages and provides a new perspective for building multilingual NLP applications.

3) Method flexibility and scalability: The proposed approach allows researchers to flexibly select and integrate external knowledge sources based on specific task requirements. Furthermore, as the external knowledge base continues to expand, the proposed approach is highly scalable and able to adapt to new tasks and challenges that may arise in the future.

References

[1] G. Lample and A. Conneau, "Cross-lingual language model pretraining," arXiv preprint arXiv:1901.07291, 2019.

[2] N. Ufer, K. T. Lui, K. Schwarz, P. Warkentin, and B. Ommer, "Multi-scale convolutions for learning context aware feature representations," arXiv preprint arXiv:1906.06978, 2019.

[3] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, June 2020.

[4] Y. Gao et al., “Retrieval-augmented generation for large language models: A survey,” arXiv preprint arXiv:2312.10997, 2023.

[5] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “Ragas: Automated evaluation of retrieval augmented generation,” arXiv preprint arXiv:2309.15217, 2023.

[6] A. J. Flanagin and M. J. Metzger, “From Encyclopaedia Britannica to Wikipedia: Generational differences in the perceived credibility of online encyclopedia information,” Inform. Commun. Soc., vol. 14, no. 3, pp. 355-374, March 2011.

[7] X. Zou, “A survey on application of knowledge graph,” J. Phys.: Conf. Ser., vol. 1487, no. 1, p. 012016, December 2020.

[8] Q. Chen, F. L. Li, G. Xu, M. Yan, J. Zhang, and Y. Zhang, “Dictbert: Dictionary description knowledge enhanced language model pre-training via contrastive learning,” arXiv preprint arXiv:2208.00635, 2022.

[9] J. Tang, R. Song, Y. Huang, S. Gao, and Z. Yu, “Semantic-aware entity alignment for low resource language knowledge graph,” Front. Comput. Sci., vol. 18, no. 4, p. 184319, December 2024.

[10] Q. Wang, Z. Mao, B. Wang, and L. Guo, “Knowledge graph embedding: A survey of approaches and applications,” IEEE Trans. Knowledge. Data Eng., vol. 29, no. 12, pp. 2724-2743, December 2017.

[11] W. Liu et al., “K-BERT: Enabling language representation with knowledge graph,” AAAI, vol. 34, no. 3, pp. 2901-2908, Septembre 2020.

[12] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” in Proc. ACL, Philadelphia, PA, USA, 2002, pp. 311-318.

[13] C. Y. Lin and E. Hovy, “Automated multi-document summarization in neats,” Proc. HLT, San Diego, CA, USA, 2002, pp. 23-27.

[14] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translationand/or Summarization, Ann Arbor, MI, USA, 2005, pp. 65-72.

[15] B. Bi, C. Wu, M. Yan, W. Wang, J. Xia, and C. Li, “Incorporating external knowledge into machine reading for generative question answering,” arXiv preprint arXiv:1909.02745, 2019.