Cross Lingual Document Grounded Conversational Question Answering

Anonymous ACL submission

Abstract

The DialDoc 2023 shared task has expanded the document-grounded dialogue task to encompass multiple languages, despite having limited annotated data. This paper assesses the effectiveness of both language-agnostic and language-aware paradigms for multilingual pre-trained transformer models in a bi-encoder-based dense retriever, concluding that the language-agnostic approach is superior. Additionally, the study investigates the impact of query rewriting techniques using large language models, such as ChatGPT, on multilingual, multi-document-grounded question-answering systems. The experiments conducted demonstrate that, for the examples examined, query rewriting does not enhance performance for rewritten queries compared to the original queries. This failure is due to topic switching in final dialogue turns and irrelevant topics being considered for query rewriting.

1 Introduction

The dominance of English as the most widely used language on the internet and for virtual assistant searches is a well-established fact [1]. However, the prevalence of English-centric content creates a language barrier for non-English speakers who wish to access information and services online. To bridge this gap, there is a growing need for multilingual knowledge-grounded question-answering dialogue systems that can enable individuals to access the internet and utilize virtual assistants in their native language. While the development of English document-grounded dialogue systems (Feng et al., 2021) has been extensively explored, the exploration of other languages remains limited.

In response to this, DialDoc 2023 shared task extends the task of document-grounded dialogue to include multiple languages with limited annotated data, such as Vietnamese and French. The development of multilingual dialogue systems poses two significant challenges: understanding queries in any language and retrieving relevant passages from a collection of documents in multiple languages and generating appropriate responses in the same language. Prior works (Clark et al., 2020; Asai et al., 2021a) in open-domain multilingual question-answering models have addressed these challenges using a retriever-reader approach. Specifically, the mDPR model, an extension of DPR (Karpukhin et al., 2020), is used to retrieve documents from a corpus, and mGEN based on mT5 (Xue et al., 2021a), generates suitable responses in the target language based on the retrieved multilingual passages. In contrast to conventional retrieval tasks, passage retrieval in conversational QA presents new challenges as each question must be interpreted within the context of the ongoing dialogue. Previous studies (Wu et al., 2022) have shown that converting a conversational question in the context of the dialogue into a standalone question can enhance the retrieval process, surpassing the performance of current state-of-the-art retrievers.

The mDPR model employs a bi-encoder architecture, utilizing a pre-trained multilingual model to independently encode the questions and passages. The encoded representations are then compared using a maximum inner product search to identify relevant passages for a given question. In this study, we evaluate two paradigms for multilingual pre-trained transformer models, namely, a language-specific paradigm and a language-agnostic paradigm, to enhance mDPR's performance. Specifically, we consider two models for multilingual sentence embedding: Language-Agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022) and mBERT (Devlin et al., 2019). LaBSE combines masked language modeling with translation language modeling to produce language-agnostic sentence embeddings, while

[1] Usage Statistics and Market Share of Content Languages for Websites, February 2023 — w3techs.com.

mBERT is a multilingual version of BERT pre-trained on a large corpus of text in multiple languages. Although both models are beneficial for multilingual sentence embeddings, in some experiments, LaBSE has demonstrated superior performance over mBERT. Additionally, we examine the impact of query rewriting techniques using large language models (LLMs) such as ChatGPT to summarize the conversational history into a more concise form and use transfer learning to generalize the performance of French and Vietnamese rewritten queries.

Therefore, in this study, we investigate the performance difference between the language-specific and language-agnostic paradigms, where we found that the language-agnostic LaBSE retriever outperforms the language-aware mBERT retriever. Additionally, we explore the impact of query rewriting on the performance of such systems. While query rewriting has been proposed as a potential solution for improving performance, our results indicate that for the considered sub-samples, rewriting queries did not lead to a significant improvement in performance.

2 Related Work

2.1 Language-agnostic Multilingual Model

The Language-agnostic BERT Sentence Embedding (LaBSE) is a method that utilizes the BERT (Devlin et al., 2019) model and a cross-lingual training technique to create sentence embeddings that can be applied to any language. By training on parallel data consisting of pairs of sentences expressing the same meaning in different languages, LaBSE is able to learn how to map sentences from different languages onto a shared high-dimensional space, where similar sentences are located close to each other and dissimilar ones are far apart. LaBSE outperforms previous state-of-the-art models in a range of cross-lingual and multilingual natural language processing tasks, including cross-lingual sentence retrieval, cross-lingual document classification, and multilingual question answering, owing to its cross-lingual training approach. This approach enables LaBSE to transfer knowledge among different languages and generate superior-quality sentence embeddings for texts in numerous languages, thereby rendering it a valuable instrument for researchers and practitioners dealing with multilingual text data.

2.2 Multilingual Query Rewriting

In this academic paper we used ChatGPT for query rewriting. ChatGPT is a GPT-based neural network model that generates high-quality responses to natural language queries. It has been shown to be effective for query rewriting and can be fine-tuned on a specific task to generate revised queries. Additionally, ChatGPT can be used for prompting, a common technique for improving model performance in natural language processing tasks. Query rewriting using prompting demonstrates the potential for improving the performance of conversational question-answering systems and enhancing information retrieval from large text corpora.

3 Dataset

DialDoc 2023 shared task dataset consists of 797 Vietnamese dialogues with a turn count of 3446 and 816 French dialogues with an average of 5 turns. These dialogues are grounded in multiple documents from nine different domains, namely Technology, Health, Health Care Services, Veterans Affairs, Insurance, Public Services, Social Security, Department of Motor Vehicles, and Student Financial Aid in the USA. Each dialogue turn in the dataset contains role annotations for the conversation between a human and a conversational agent, with the latest turn presented at the beginning of the dialogue. The retrieval dataset includes query dialogue data, positive passages, and negative passages. Positive passages contain the answer to the given query and are within the document, while negative passages do not contain the answer to the query in focus and are also within the document.

4 Methodology and Experiments

The prevailing paradigm for document-grounded question-answering models involves a retriever-reader approach that comprises of a document retrieval module, a reranker module, and an answer generation module. However, in this study, our main focus has been on the multilingual retriever component, while we have utilized a fixed XLM-R reranker and a multilingual T5 model as the generator (Xue et al., 2021b).

4.1 Retrieval: Language Agnostic vs Language-aware

In this paper, we employ the multilingual dense passage retriever (m-DPR) (Asai et al., 2021b) to

Model	Pretrained on	Trained on	Eval On	R@1	R@5	R@10	R@20
LaBSE	Chinese + English	French + Vietnamese	French + Vietnamese	**0.646**	**0.816**	**0.864**	**0.9**
LaBSE	Chinese + English	French + Vietnamese	French	0.575	0.765	0.823	0.869
LaBSE	Chinese + English	French + Vietnamese	Vietnamese	0.752	0.894	0.925	0.954
XLMR	Chinese + English	French + Vietnamese	French + Vietnamese	0.547	0.747	0.796	0.839
XLMR	Chinese + English	French + Vietnamese	French	0.446	0.67	0.725	0.779
XLMR	Chinese + English	French + Vietnamese	Vietnamese	0.654	0.83	0.871	0.903

Table 1: Comparison of language agnostic vs multilingual dense passage retrieval approaches.

retrieve passages from multilingual document collections. However, the bi-encoders used in m-DPR consist of two different models: LaBSE, which is designed for the language-agnostic paradigm, and XLM-R, which is suitable for the language-aware setting. To prepare the models, we first pre-train them using the English and Chinese portions of a document-grounded dataset. We then fine-tune the models on three different combinations of target datasets, namely French and Vietnamese, French only, and Vietnamese only. Finally, we evaluate the performance of the models on the corresponding validation portions of each dataset combination.

To train the mDPR models using the English and Chinese splits, a document corpus is required to retrieve passages. We created our corpora by concatenating the gold passages given for each query and subsequently using BM25 to retrieve the positive and negative passages for each instance. The dataset was then divided using a fixed seed and test ratio of 20%, while the number of training epochs remained constant at 50.

4.2 Zero-Shot Multilingual Query Re-writing

To improve the efficiency of the retriever module, we postulated that converting the query and dialogue context history into more concise and informative questions would be advantageous. Drawing inspiration from the accomplishments of large language models, we utilized ChatGPT for query rewriting. A specific prompt structure was employed for the ChatGPT model, where the question was generated from the last turn in the query, and the context encompassed all preceding turns concatenated in reverse order. The template of the prompt that we employed is provided below:

Rewrite the question into an informative query explicitly mentioning relevant details from the provided context. Context : {context} Question : {question} Re-written Question :

Our study's outcomes, which compared language-agnostic and multilingual paradigms, demonstrated that LaBSE-based retrievers outperformed other methods for multilingual retrieval tasks. As a result, we opted to utilize the LaBSE-based mDPR retriever module for all subsequent experiments. We also evaluated the impact of utilizing forward-order context, but the results indicated that it accentuated irrelevant information.

We employed the gpt-3.5-turbo model, with a temperature of 0.2, for query rewriting by translating the English prompt into Vietnamese and French. This permitted a broader spectrum of language contexts to be covered. However, due to resource constraints, we could only utilize ChatGPT to rewrite a limited number of English queries and the complete French and Vietnamese validation portions. To compensate, we generated a new dataset that included rewritten queries for 2000 English queries and dialogue histories that were randomly chosen.

5 Results and Discussion

Language agnostic retrievers outperform language-aware retrievers. In Table 1, we present the results of our experiments, where we first pre-trained the retriever models LaBSE and XLM-R on Chinese + English data, and then fine-tuned them on various combinations of French and Vietnamese document grounded datasets, as described in Section 4.1. The findings demonstrate that the LaBSE-based mDPR retriever model performed better than the XLM-R-based mDPR retriever model, in terms of all metrics and training dataset combinations. Although XLM-R, which is based on RoBERTa and has 125M model parameters, was trained on unsupervised cross-lingual data, LaBSE still outperformed it.

Multilingual Query Rewriting does not lead to better performance. The results presented in Table 2 provide evidence that the incorporation of multilingual query rewriting does not lead to enhanced performance. More precisely, the LaBSE

Model	Trained on	Eval On	R@1	R@5	R@10	R@20
LaBSE	English (Raw)	French (Raw)	0.454	0.696	0.776	0.836
LaBSE	English (Raw)	Vietnamese (Raw)	0.51	0.776	0.838	0.894
LaBSE	English (ChatGPT Rewritten)	French (ChatGPT Rewritten)	0.162	0.328	0.382	0.464
LaBSE	English (ChatGPT Rewritten)	Vietnamese (ChatGPT Rewritten)	0.264	0.486	0.576	0.652

Table 2: Comparison of transfer learning approaches using different query-rewriting approaches. Raw Query refers to Current turn + History.

model, which was trained on unmodified subsets of English data, demonstrated superior knowledge transfer abilities compared to models trained on queries that were rewritten by ChatGPT. Thus, further research is necessary to elucidate the reasons for this suboptimal performance.

5.1 Error Analysis

Original vietnamese query:
Dialogue History:
user: thiết bị tiêu dùng khác để làm gì?
agent: để khuyến khích lối sống lành mạnh, chẳng hạn như, quy mô kết nối hoặc máy theo dõi tim mạch, cũng là một khả năng của IoT.
user: Thiết bị IoT cũng có thể làm gì?
agent: có thể được sử dụng để kiểm soát cơ sở hạ tầng quan trọng như cầu để cung cấp truy cập vào tàu.
user: các ứng dụng của IoT có giới hạn trong các lĩnh vực này không?
agent: các ứng dụng của IoT không chỉ giới hạn trong các lĩnh vực này.

Original Query:
ARM đã "nhanh chân" trong việc nhận ra rằng, ổ đĩa có xu hướng sử dụng các bộ vi điều khiển 32-bit là giải pháp cho những người như thế nào?

Rewritten Query:
Trong lĩnh vực IoT, các ứng dụng có giới hạn trong những lĩnh vực nào và thiết bị IoT có thể được sử dụng để làm gì, bao gồm cả việc khuyến khích lối sống lành mạnh và thiết bị tiêu dùng khác?

Translated english query:

Dialogue History
user: other consumer devices for what?
agent: to promote healthy lifestyles, for example, connected scales or heart monitors, is also an IoT possibility.
user: What can IoT devices do too?
agent: can be used to control critical infrastructure such as bridges to provide access to ships.
user: are IoT applications limited to these areas?
agent: applications of IoT are not limited to these areas.

Question:
ARM was "quick feet" in realizing that the drive tends to use 32-bit microcontrollers is the solution for those?

Rewritten Query:
In the field of IoT, what areas are the applications limited to and what can IoT devices be used for, including promoting healthy lifestyles and other consumer devices?

Figure 1: Query Rewriting in the case of Topic Switching in the last dialogue turn

This study focuses on the evaluation of the performance of rewritten queries generated by ChatGPT in comparison to the original queries consisting of a question and context. Moreover, a comprehensive error analysis is conducted to identify the gaps in the rewritten queries. Figure 1 presents notable observations. The findings reveal that the quality of the rewritten queries generated by ChatGPT is inferior to that of the original queries. Further investigation shows that topic switching often occurs in the last turn of the conversation, resulting in the generation of rewritten queries that incorporate non-relevant context. This phenomenon is illustrated in Figure 1. The switching of topics adversely affects the relevance and accuracy of the rewritten queries. Additionally, the rewriting process tends to summarize both relevant and non-relevant topics from the conversation, as shown in Figure 2. This approach results in a lack of specificity and clarity in the rewritten queries, further impeding their quality and effectiveness. Furthermore, the prompts are created manually by visually inspecting the generated outputs. While this method allows for quality control of the prompts, it is inherently subjective and prone to human biases. Thus, it is essential to explore advanced prompting methods to enhance the overall quality of the rewritten queries, as suggested in (Prompt, 2021).

6 Conclusion

This paper investigates the effectiveness of language-agnostic and language-aware paradigms for multilingual pre-trained transformer models in a bi-encoder-based dense retriever, in the context of the DialDoc 2023 shared task. The paper also evaluates the impact of query rewriting on system performance. Our findings indicate that the language-agnostic approach outperforms the language-aware approach. However, for the considered subsamples, query rewriting does not improve the performance of the original queries. Furthermore, the observed topic switching in the last conversation turns and ChatGPT's tendency to summarize non-relevant topics lead to a lower accuracy and relevance of the rewritten queries compared to the original queries.

7 Limitations and Potential Risks

The limitations of this study are primarily due to budget and credit constraints. Consequently, our observations are based on a sample size of 2000, leading to limited generalizability of findings. The study is also subject to the risk of "hallucinations" in ChatGPT's responses, which may lead to imprecision in query rewriting. The study suggests further investigation into these issues to improve the accuracy and reliability of the results. We recommend further investigation into these limitations,

as well as any potential societal biases present in our dataset, to enhance the reliability and performance of query rewriting.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Zeqiu Wu, Yi Luan, Hannah Rashkin, David Reitter, Hannaneh Hajishirzi, Mari Ostendorf, and Gaurav Singh Tomar. 2022. CONQRR: Conversational query rewriting for retrieval with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10000–10014, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021a. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mt5: A massively multilingual pre-trained text-to-text transformer.

A Example Appendix

In this appendix, we present cases containing failure of query writing in terms of performance boost as compared to the original query.

A.1 Query Rewriting considering non-relevant topics

One of the examples where Query Rewriting was seen to consider non-relevant topics can be seen in Figure 2.

Original french query:

Dialogue History

user: Qu'est-ce qui est fait avec une partie de l'énergie?
agent: La climatisation, le chauffage, l'éclairage et autres.
user: Quels facteurs limitent l'efficacité des trains à sustentation magnétique classiques?
agent: La traînée aérodynamique.
agent: Se déplacer linéairement le long de la voie. <user> Qu'est-ce que le sujet peut faire le long de la voie?

Original Query:
Comment peut-on entraîner un train sans disposer d'une plaque de réaction séparée ?

Rewritten Query:
Quels facteurs limitent l'efficacité des trains à sustentation magnétique classiques, tels que la climatisation, le chauffage, l'éclairage et autres, et que fait-on avec une partie de l'énergie, sachant que l'agent se déplace linéairement le long de la voie et que la traînée aérodynamique est utilisée pour l'entraînement du train sans disposer d'une plaque de réaction séparée ?

Translated query:

Dialog History:
user: What is done with part of the energy?
agent: Air conditioning, heating, lighting and others.
user: What factors limit the efficiency of conventional magnetic levitation trains?
agent: Aerodynamic drag.
agent: Move linearly along the lane. <user> What can the subject do along the way?

Original Query:
How can you drive a train without having a separate reaction plate?

Rewritten Query:
What factors limit the efficiency of conventional magnetic levitation trains, such as air conditioning, heating, lighting and others, and what do we do with some of the energy, knowing that the agent moves linearly along of the track and that the aerodynamic drag is used for driving the train without having a separate reaction plate?

Figure 2: Query Writing considering non-relevant topics