Language-Agnostic Transformers and Assessing ChatGPT-Based Query Rewriting for Multilingual Document-Grounded QA

Srinivas Gowriraj* , **Soham Dinesh Tiwari***, **Mitali Potnis***, **Srijan Bansal**, **Teruko Mitamura**, and **Eric Nyberg**
{sgowrira, sohamdit, mpotnis, srijanb, teruko, en09}@andrew.cmu.edu
Language Technologies Institute, Carnegie Mellon University

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 passage retriever (DPR), 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, document-grounded question-answering systems. The experiments conducted demonstrate that, for the examples examined, query rewriting does not enhance performance 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

English dominates as the most widely used language on the internet, and for communicating with virtual assistants [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: (i) understanding queries in any language and retrieving relevant passages from a collection of documents in multiple languages (ii) 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 multilingual DPR (mDPR) model, an extension of DPR (Karpukhin et al., 2020), is used to retrieve documents from a corpus. A multilingual reader based on multilingual T5 (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 question answering (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 rewriting the question using the dialogue context 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 encode the questions and passages independently. 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 as mDPR bi-encoders, namely, a language-agnostic paradigm and a language-aware paradigm. Specifically, we consider two models for multilingual sentence embedding: Language-Agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020). LaBSE combines masked language modeling with translation language modeling to produce language-

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

*These authors contributed equally to this work.

agnostic sentence embeddings, while XLM-R is a cross-lingual version of RoBERTa (Liu et al., 2019) pre-trained on a large corpus of text in over 100 languages using a self-supervised approach. Although both models are beneficial for multilingual sentence embeddings, based on our experiments, it has been observed that LaBSE outperforms XLM-R. Additionally, we examine the impact of query rewriting techniques using large language models (LLMs) such as ChatGPT to summarize the conversational history more concisely and use transfer learning to generalize to French and Vietnamese rewritten queries.

Therefore, in this study, we investigate the performance difference between the language-aware and language-agnostic paradigms, where we found that the language-agnostic LaBSE retriever outperforms the language-aware XLM-R 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 rewriting queries did not significantly improve performance for the considered sub-samples. Our code is available on GitHub [2].

2 Related Work

2.1 Language-agnostic Multilingual Model

Language-agnostic BERT Sentence Embedding (LaBSE) model is essentially the BERT (Devlin et al., 2019) model trained with a cross-lingual training technique to create language agnostic sentence embeddings for many languages. 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. Language agnosticism enables LaBSE to transfer knowledge across different languages and generate superior-quality sentence embeddings for texts in numerous languages, thereby making it a valuable instrument for researchers and practitioners dealing with multilingual text data. We have employed LaBSE in our work due to its shared embedding space and its ability to capture contextual information across multiple languages enabling strong cross-lingual performance and knowledge transfer across multiple languages.

2.2 Multilingual Query Rewriting

GPT models, including ChatGPT, possess a remarkable capability for comprehending and interpreting natural language (Haleem et al., 2023; Walid, 2023). ChatGPT has proven to be highly capable of high-quality responses to natural language queries. Additionally, it is also effective at rewriting long contextual information into compact queries (Wang et al., 2023). Prompting methods (White et al., 2023; Zuccon and Koopman, 2023) are used to steer the LLM's behaviour for desired outcomes without updating model weights. In this academic paper, we have used ChatGPT for the purpose of query rewriting. Query rewriting by prompting ChatGPT has potential to improve the effectiveness of conversational question-answering systems and aiding the retrieval of information from extensive text collections.

3 Dataset

DialDoc 2023 shared task dataset consists of 797 Vietnamese dialogues with an average turn count of 4 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 turns in reverse chronological order, the latest turn first in dialogue history. 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 are closely related to the document but they do not contain the answer to the query in focus.

4 Methodology and Experiments

The prevailing paradigm for document-grounded question-answering models involves a retriever-

[2] https://github.com/srinivas-gowriraj/Multilingual_QA/

Model	Pretrained	Finetuned	Evaluated	R@1	R@5	R@10	R@20
LaBSE	zh + en	fr + vi	fr + vi	**0.65**	**0.82**	**0.86**	**0.90**
LaBSE	zh + en	fr + vi	fr	0.57	0.76	0.82	0.87
LaBSE	zh + en	fr + vi	vi	0.75	0.89	0.92	0.95
XLMR	zh + en	fr + vi	fr + vi	0.55	0.75	0.80	0.84
XLMR	zh + en	fr + vi	fr	0.45	0.67	0.72	0.78
XLMR	zh + en	fr + vi	vi	0.65	0.83	0.87	0.90

Table 1: Performance comparison of language-agnostic versus language-aware multilingual dense passage retrieval approaches pre trained on Chinese (zh) + English (en) and fine-tuned on French (fr) + Vietnamese (vi).

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 fixing XLM-R as reranker. However, we did experiment with the Fusion-in-Decoder (FiD) approach (Raffel et al., 2020) to modify the mT5 model previously being used as the answer generator (Xue et al., 2021b).

4.1 Retrieval: Language Agnostic vs Language-aware

In this paper, we employ the multilingual dense passage retriever (mDPR) (Asai et al., 2021b) to retrieve passages from multilingual document collections. However, the bi-encoders used in mDPR 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 sets of each dataset combination.

The mDPR models are trained using the English and Chinese splits, employing gold passages along with 10 hard negatives mined through BM25. [3]. The dataset is divided into an 80-20 train-test split using a consistent seed. The training process for all configurations persists for 50 epochs.

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 rewritten using 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 : {dialogue history} Question : {last-turn} 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.

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 (zh) + English (en) data, and then fine-tuned them on various combinations of French (fr) and Vietnamese (vi) document grounded datasets, as described in Section 4.1. The findings demonstrate that the LaBSE-based mDPR retriever model outperformed the XLM-R-based mDPR retriever model, in all metrics and training dataset combinations. Although XLM-R, which is based on RoBERTa (a more advanced version of BERT) and has 125M model parameters, was trained on unsupervised cross-lingual data, LaBSE still outperformed it. The BERT-based architecture has 110M trainable parameters.

[3] https://www.analyticsvidhya.com/blog/2021/05/build-your-own-nlp-based-search-engine-using-bm25.

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 for the tested examples. More precisely, the LaBSE model, which was trained on unmodified subsets of English (en) 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.

Trained on	Eval On	R@1	R@5	R@10	R@20
en (Raw)	fr (Raw)	0.45	0.70	0.77	0.84
en (Raw)	vi (Raw)	0.51	0.78	0.84	0.89
en (ChatGPT)	fr (ChatGPT)	0.16	0.33	0.38	0.46
en (ChatGPT)	vi (ChatGPT)	0.26	0.49	0.58	0.65

Table 2: Comparison of transfer learning approaches using different query-rewriting approaches. **Raw** refers to the original dialogue query i.e. current turn + History while **ChatGPT** refers to the query re-written by ChatGPT using current turn and dialogue history. We use LaBSE-based mDPR for all the above settings.

5.1 Error Analysis of Rewritten Queries

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 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, and hallucinate information, as shown in Figure 2. This approach lacks 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 vulnerable to human biases. Thus, it is essential to explore advanced prompting methods to enhance the overall quality of the rewritten queries, as suggested in (Liu et al., 2023).

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. The paper also evaluates the impact of query rewriting on task performance. Our findings indicate that the language-agnostic approach outperforms the language-aware approach. However, for the considered subsamples, query rewriting did not improve the performance over the original queries. Furthermore, the observed topic switching in the conversations' last turns, and ChatGPT's tendency to summarize non-relevant topics and hallucination lead to less accurate rewritten queries compared to the original queries.

7 Limitations, Potential Risks, and Future Work

The limitations of this study are primarily due to budget and credit constraints. Consequently, our query rewriting observations are based on a sample size of 2000, leading to limited generalizability of findings. Another limitation of the limited resources was the limited context size of ChatGPT and the relatively long nature of the questions in our dataset. Hence, we could not test prompting ChatGPT with in-context examples for better query rewriting performance. Hence one potential future work is testing the performance of query rewriting using in-context examples. Finally, ChatGPT architecture is not open source, preventing us from testing advanced prompting methods. Hence another future work would be to test query rewriting using open source models and with advanced fine-tuning methods like Prefix Tuning (Li and Liang, 2021) and Prompt Tuning (Lester et al., 2021).

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

Original vietnamese query	Translated english 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?	**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. **Original Query (translated):** ARM was "quick feet" in realizing that the drive tends to use 32-bit microcontrollers is the solution for those? **Rewritten Query (translated):** 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: An erroneous rewritten query (ChatGPT) occurred wherein the subject matter abruptly changed during the final exchange of dialogue, as highlighted in yellow.

Original french query	Translated english 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 ?	**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: Erroneous rewritten query (ChatGPT) which considered non-relevant topics and hallucinated information

References

Sumit Agarwal, Suraj Tripathi, Teruko Mitamura, and Carolyn Penstein Rose. 2022. Zero-shot cross-lingual open domain question answering. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 91–99, Seattle, USA. Association for Computational Linguistics.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

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.

Abid Haleem, Mohd Javaid, and Ravi Singh. 2023. An era of chatgpt as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2:100089.

Gautier Izacard and Edouard Grave. 2021a. Leveraging passage retrieval with generative models for open domain question answering.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. 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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. 55(9).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Siamak Shakeri, Noah Constant, Mihir Sanjay Kale, and Linting Xue. 2021. Towards zero-shot multilingual synthetic question and answer generation for cross-lingual reading comprehension.

Hariri Walid. 2023. Unlocking the potential of chatgpt: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023. Can chatgpt write a good boolean query for systematic review literature search?

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt.

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.

Guido Zuccon and Bevan Koopman. 2023. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness.

A Appendix

A.1 Reader: mT5 and Fusion-in-Decoder

In recent years, the use of the multilingual Text-to-Text Transfer Transformer (mT5) (Xue et al., 2021b), a multilingual variant of Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) has gained popularity in multilingual question answering tasks (Shakeri et al., 2021). mT5 has the ability to learn the representations of text that capture the nuances of language across different languages and contexts, allowing it to excel in multilingual settings. To leverage this growing popularity of the mT5 generative reader model, our work involves employing a Fusion-in-Decoder (FiD) (Izacard and Grave, 2021a) as a reader in combination with mT5. FiD in combination with mT5 has been proven to boost the performance of answer extraction as compared to using mT5 alone (Agarwal et al., 2022) by encoding the reranked passages individually one-by-one and concatenating them together while passing them for the decoder stage.

We conducted further analysis on the impact of our language-agnostic multilingual retriever selection by employing a two-step process. Firstly, we retrieved relevant passages using the chosen retriever, and subsequently, we utilized the mT5 reader to generate responses based on these retrieved passages. Two different versions of the reader were employed: the vanilla mT5 and the mT5 with Fusion-in-Decoder (FiD) (Izacard and Grave, 2021b). The vanilla mT5 reader takes the query concatenated with the retrieved passages as its input. On the other hand, the FiD-mT5 reader independently encodes each retrieved passage along with the query. The encoded representations are then concatenated and passed to the decoder. As a result, the evidence fusion takes place solely within the decoder. To assess the quality of the generated responses, we employed BLEU, Rouge-L, and F1 metric scores.

We further evaluate the impact of our choice of language-agnostic multilingual retriever by passing retrieved passages to mT5 reader to generate response. We used two different readers which are mT5 (vanilla) and mT5 with Fusion-in-Decoder (FiD) (Izacard and Grave, 2021b). Vanilla mT5 takes query concatentate with retrieved passages as input while FiD-mT5 encodes each retrieved passage along with query independently which is then concatenated and passed to the decoder. The model thus performs evidence fusion in the decoder only. We evaluate the generated response using BLEU, Rouge-L, and F1 metric scores.

Model	Pre-trained	Evaluated	F1	BLEU	ROUGE-L
mT5	fr + vi	fr + vi	62.43	40.87	62.96
mT5 + FiD	fr + vi	fr + vi	64.83	42.22	64.73
mT5	fr	fr	59.89	39.42	58.55
mT5 + FiD	fr	fr	56.76	41.47	60.00
mT5	vi	vi	65.34	45.93	63.03
mT5 + FiD	vi	vi	68.22	45.72	65.61

Table 3: Reader performance comparison of vanilla mT5 versus mT5 in combination with FiD

Fusion-in-Decoder (FiD) improves the overall reader performance. As illustrated in Table 3 we observed an improvement of approximately 4.2%, 3.51%, and 2.80% in F1, BLEU, and Rouge-L scores, respectively when we include Fusion-in-Decoder along with mT5 in the case of both Vietnamese (vi) and French (fr) languages thus proving that Fusion-in-Decoder does indeed help in bolstering our Reader performance.