VISHC SUMMER SCHOOL.

BILINGUAL DOCUMENT-BASED QUESTION ANSWERING FOR MEDICAL DOMAIN

Students:
Dang Dinh Dang Khoa
Vu Hong Phuc

Supervisor
Prof. Le Duy Dung

Contributors:
Vo Diep Nhu
Vuong Duong

Abstract

This paper explores the application of Retrieval-Augmented Generation (RAG) pipeline in the medical domain, introducing a new embedding model to enhance cross-lingual English-Vietnamese (EN-VI) performance. The study evaluates the efficacy of diverse models and methodologies within the RAG pipeline to improve information retrieval and text generation in the medical context.

1 Introduction

In recent years, Large pre-trained language models (PLMs), such as T5 and GPT3, have significantly advanced the area of natural language processing (NLP) by displaying astounding performance on a variety of downstream tasks [7]. These PLMs can predict the results on downstream tasks without access to any external memory or raw text, as a parameterized implicit knowledge base [11], because they have learnt a significant amount of in-depth information from the pre-training corpus [6]. However, these PLMs cannot "*easily expand or revise their memory [...] and may provide hallucinations*" [2]. The solution that seems obvious at first glance is to augment the input of PLMs with external information (such as encyclopedias and books).

Retrieval-Augmented Generation (RAG): A new learning paradigm that combines PLMs with conventional IR techniques. [2]. RAG has attained state-of-the-art performance in several knowledge-intensive NLP tasks. [5] The RAG model offers numerous advantages over its large-scale PLM competitors: (i) The information is explicitly gained rather than implicitly retained in model parameters, allowing for tremendous scalability; (ii) The model generates outputs based on some retrieved references rather than starting from scratch, which lessens the challenge of hallucination from text generation.

Bilingual Sentence Embedding for Medical Texts: To perform well on downstream NLP tasks (e.g information retrieval in our study), we need a good text embedding model. [9] Our focus is on cross-lingual English-Vietnamese (EN-VI) information retrieval and question-answering. Thus, we need a Vietnamese sentence embedding model for medical domain. We finetune a generic multilingual sentence embedding model to learn specific medical representation with an approach called knowledge distillation. [10] We then use it for embedding the medical documents in our RAG pipeline.

Our contributions are twofold. First, we apply the knowledge distillation method to train a bilingual (en-vi) sentence embedding on medical domain. Second, we develop a retrieval-augmented generation pipeline with a chatbot demo for medical documents to solve cross-lingual medical question-answering task.

2 Methodology

2.1 Retrieval-Augmented Generation Pipeline

In this section, we present the methodology for the Retrieval-Augmented Generation (RAG) Pipeline, outlining the key steps involved: embedding, retrieval and generation. The goal of the RAG Pipeline is to efficiently retrieve relevant documents or passages and then generate human-readable responses or content based on the retrieved information.

Before the question-answering process begins, the following data preprocessing steps are performed:

Document Embedding: The documents, typi-

cally consisting of sentence-to-paragraph-sized text, are embedded using a pre-trained embedding model. This embedding process converts the text into vector representations, enabling efficient storage and retrieval.

Vector Database: The embedded documents are indexed and stored in a vector database, in this case, Weaviate. This database allows for efficient retrieval of context vectors during the retrieval phase. hyperref

The retrieval phase is when relevant documents or contexts are retrieved from the indexed corpus based on a user's query or question. This phase involves the following steps:

Query Vector Creation: Given a user's question or query, the retriever creates a sparse or dense vector representation known as the query vector. This vector encodes the semantic information of the question.

Vector Comparison: The query vector is compared against all indexed context vectors in the database. Various similarity metrics, such as cosine similarity or semantic similarity, can be employed for this comparison. The aim is to identify the k most similar context vectors, where k is a user-defined parameter.

Context Retrieval: The k most similar context vectors are retrieved from the vector database. These contexts serve as the basis for generating responses or answers to the user's query.

Following the retrieval phase, the generation phase takes the retrieved context and the user's original question to generate coherent and contextually relevant responses. This phase includes the following steps:

Question and Context Input: The retrieved context vectors are passed to the reader model along with the original user's question. The combination of the question and context provides essential information for generating accurate responses.

Answer Generation: The RAG Pipeline generates the answer by extracting the corresponding text span from the retrieved context. The answer can be generated using a language model that takes the question and context as input.

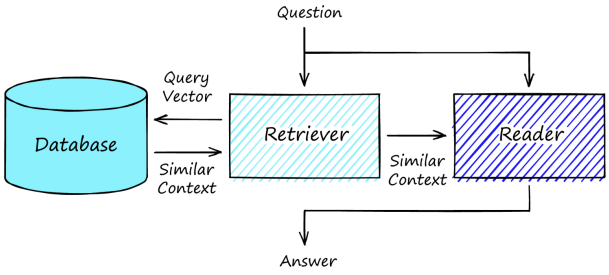

Figure 1: RAG pipeline

2.2 Bilingual Sentence Embedding with Knowledge Distillation

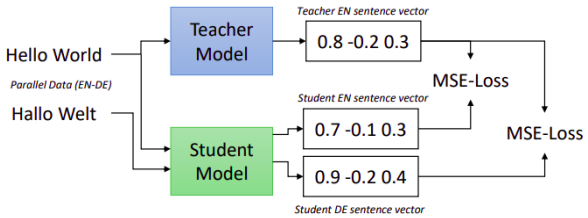

Figure 2: "Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector" [10]

In this section, we present the knowledge distillation training process. This method requires a monolingual teacher model M, which maps a sentence in one source language s to a dense vector space and specializes in a particular domain. Then, we need parallel (translated) sentences $((s_1, t_1), \ldots, (s_n, t_n))$ with s_i being a sentence in source language and t_i a sentence in target language. [10] We train a multilingual general-domain student model $\hat{M}$ such that $\hat{M}(s_i) \approx M(s_i)$ and $\hat{M}(t_i) \approx M(s_i)$. For a given mini-batch B, we minimize the mean-squared loss:

$$\frac{1}{|B|} \sum_{j \in B} \left((M(s_j) - \hat{M}(s_j))^2 + (M(s_j) - \hat{M}(t_j))^2 \right)$$

From this setting, the student model $\hat{M}$ learns the representation of the teacher model M. In our experiments, we use `BioSimCSE` [1] - an English SBERT model pre-trained on biomedical texts - as teacher model M and use `paraphrase-multilingual-mpnet-base-v2` [9] - a

multilingual SBERT model pre-trained on general domain texts - as student model $\hat{M}$.

We trained for a maximum of 5 epochs with batch size 64; 10,000 warm-up steps; and a learning rate of 2e-5. As development set, we measured the mean-squared loss on hold-out parallel sentences. We will denote the finetuned model as `vishc-med-qa`. This is the **link** to our model hosted on Huggingface Hub.

3 Experiments

3.1 Datasets

In our experiments, we used the following datasets:

- **VinUni EVMed** [12]
  - Manually translated EN-VI sentence pairs collected from the abstract of many medical documents crawled from the Internet.
  - Includes 358,885 EN-VI sentence-alignment pairs of medical text
  - An example parallel sentence:
    - EN: Partial embolization of nidus
    - VI: Nút tắc một phần ổ dị dạng
- **ViHealthQA** [3]
  - A Vietnamese dataset created from user-submitted queries on health-related websites and expert-submitted responses.
  - Includes 10,000 VI medical text corpus with 2,000 queries
  - An example query-corpus:
    - Query: Có những loại siêu âm thai nào?
    - Corpus: Về cơ bản có 7 kiểu siêu âm ...

3.2 Embeddings

3.3 Retriever

Our experimentation relies on ViHealthQA dataset. We evaluate two distinct embedding models in our document retrieval experiment: `paraphrase-multilingual-mpnet-base-v2` and our fine-tuned model `vishc-med-qa`. For efficient and scalable document retrieval, we utilize the Weaviate vector database.

3.3.1 Search Methods

In our document retrieval experiment, we employ three distinct search methods:

Similarity Search: This method involves calculating the cosine similarity between the query vector and document vectors in the Weaviate database. Documents are ranked based on their similarity to the query, with highly similar documents receiving higher ranks.

BM25 Search: We implement the BM25 search method, a popular ranking algorithm for information retrieval tasks. BM25 considers term frequencies and document lengths to calculate relevance scores. Documents are ranked according to their BM25 scores.

Hybrid Search: The hybrid search method combines elements of similarity search and BM25 search. It leverages a weighted combination of both similarity and BM25 scores to produce the final ranking of documents. We use two variants of hybrid search with weighting factors set to 0.5 and 0.75 to explore the impact of different weightings on retrieval performance.

3.3.2 Evaluation Metrics

In this section, we introduce two crucial metrics employed for retriever evaluation: F1 Score and Mean Reciprocal Rank (MRR)

F1 Score: The F1 Score is particularly relevant in Document Question Answering (DQA) because it elucidates the retriever's ability to provide both precise and comprehensive document selections. In this research, we employ F1 Score to ensure that the retriever is not only retrieving relevant documents but doing so accurately and comprehensively.

Mean Reciprocal Rank (MRR): MRR is particularly apt for assessing the retriever's efficiency in presenting relevant documents to users in ranked order, which aligns seamlessly with the DQA setting where document ranking plays a pivotal role. MRR measures the average of the reciprocal ranks of the first relevant document retrieved for a set of queries. It encapsulates the idea that the most relevant documents should ideally appear at the top of the retrieval results. Thus, a higher MRR score signifies a retriever that excels in quickly delivering pertinent documents to users.

3.4 Translation

In this section, we outline the experimental setup for our machine translation task, which aims to assess the translation quality of three distinct models: vinai-translate-en2vi, gpt-3.5-turbo, and gpt-3.5-turbo (with context). We introduce the BLEU (Bilingual Evaluation Understudy) score as the evaluation metric, discuss the prompt engineering strategy used for the context-aware model, and detail the dataset employed for our experiments.

3.4.1 Dataset

Our experiment utilizes a specialized "Parallel Paragraph Medical Corpus" dataset. This corpus comprises 50 paragraphs of medical content, with each paragraph accompanied by its English-to-Vietnamese translation.

3.4.2 BLEU Score

The BLEU score is a widely accepted metric for evaluating the quality of machine-generated translations in natural language processing tasks. It measures the similarity between machine-generated translations and reference human translations by analyzing n-grams (subsequences of n words) and considering precision and brevity penalties. The BLEU score is particularly effective for assessing the fluency and adequacy of machine translations and produces scores between 0 and 1, with higher scores indicating better translation quality. In our experiments, we calculate the mean BLEU score over 50 data points to provide a comprehensive evaluation of the models' performance.

3.4.3 Models

We evaluate three machine translation models in our experiment: vinai-translate-en2vi [4], gpt-3.5-turbo, and gpt-3.5-turbo (with context).

gpt-3.5-turbo (with context): leverages a context-aware approach to improve translation quality. We employ prompt engineering techniques to provide the model with contextual information by passing a Vietnamese paragraph as context in the prompt. This enables the model to generate translations that are more contextually coherent.

4 Result and Discussion

4.1 Semantic Textual Similarity

The goal of semantic textual similarity (STS) is to assign a pair of sentences a score indicating their semantic similarity. In our experiment, after generating the sentence embeddings for each sentences pair, we compute cosine similarity as the score function, as recommended by [8]. We compare the performance of the finetuned model with the pre-trained student model on the `test` split of the VinUni EVMed dataset. In this split, we already have annotated similar EN-VI pairs as positive pairs, and we generate negative pairs from random shuffling.

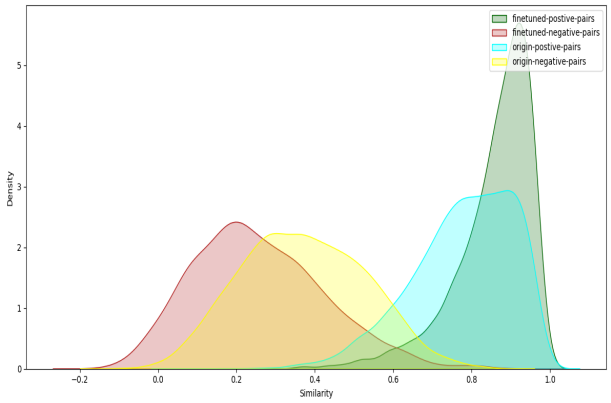

Figure 3: Comparison of kernel density distribution between 2 models

We visualize the similarity distribution (with Gaussian kernel) of positive and negative pairs computed by the 2 models. From the plot, we observe that the finetuned model is more confident in determining positive and negative pairs than the pre-trained student model, indicating much better aligned vector spaces between EN and VI medical texts.

4.2 Retrieval

In the evaluation of retriever component, the choice of appropriate evaluation metrics is of paramount importance. These metrics serve as the compass by which we gauge the system's proficiency in retrieving relevant documents, thereby influencing the overall system's effectiveness. In this section, we introduce two crucial metrics employed for retriever evaluation: F1 Score and Mean Reciprocal Rank (MRR).

Table 1: Retriever evaluation on ViHealthQA dataset using hybrid search with alpha are 0.5 and 0.75

Dataset	Model	en-vi				vi-vi			
		F1		MMR		F1		MMR	
		0.5	0.75	0.5	0.75	0.5	0.75	0.5	0.75
ViHealthQA	paraphrase-multilingual-mpnet-base-v2	0.2834	0.3413	0.1746	0.1787	0.4671	0.4343	0.1740	0.1784
	vishc-med-qa	0.2606	0.312	0.1602	0.1648	0.4275	0.4039	0.1602	0.1648

We evaluate 2 embedding models in retrieval context on ViHealthQA dataset through F1 score and Mean reciprocal rank (MMR) using hybrid search with alpha are 0.5 and 0.75. The evaluation is conducted on two types of queries: English queries and Vietnamese queries. The results are displayed in Table 1.

Analyzing the performance of both models on English queries reveals that both models perform better on Vietnamese query than on English query given that Vietnamese documents.

While the base model achieved F1 Scores of 0.2834 and 0.1746 and MMR scores of 0.1746 and 0.1787 for $\alpha = 0.5$ and $\alpha = 0.75$, respectively, the fine-tuned model obtained F1 scores of 0.2606 and 0.3120 and MMR scores of 0.1620 and 0.1648 for $\alpha = 0.5$ and $\alpha = 0.75$, respectively. These results indicate that the based model performs relatively well on English query, compared to the fine-tuned model.

The similar trend can be observed in the Vietnamese queries context. While the base model achieved F1 Scores of 0.4671 and 0.4343 and MMR scores of 0.1740 and 0.1784 for $\alpha = 0.5$ and $\alpha = 0.75$, respectively, the fine-tuned model obtained F1 scores of 0.4275 and 0.4039 and MMR scores of 0.1602 and 0.1648 for $\alpha = 0.5$ and $\alpha = 0.75$, respectively. These results also indicate that the based model performs relatively well on Vietnamese query also, compared to the fine-tuned model.

This is an intriguing finding that takes us by surprise. Because of the fixed timeline of the VISHC Summer School, we do not have enough time to investigate the rationale for the reduction in retrieval performance after finetuning. Our initial intuitions suggest 2 possible reasons: (i) The model has overfitted the training dataset, making it perform worse on other unseen data; (ii) The length of the sentences pairs used for training is relatively short, while we mostly evaluate the model on Vietnamese documents (whose lengths are much longer than the aforementioned sentences), thus, we suspect a difference between performance of a sentence embedding model and a document embedding one.

4.3 Translation

In this section, we provide a comprehensive evaluation of our machine translation experiment, focusing on the performance of three distinct models: vinai-translate-en2vi, gpt-3.5-turbo, and gpt-3.5-turbo with context. We assess their translation quality using the widely recognized BLEU score, a metric designed to quantify the quality of machine-generated translations by comparing them to human references.

Results from Table 2 show that vinai-translate-en2vi, gpt-3.5-turbo, and gpt-3.5-turbo (with context) achieved BLEU scores of 0.2814, 0.2951, and 0.4049, respectively. It is notable that gpt-3.5-turbo (with context) achieved the highest BLEU score. Notably, the introduction of context, as seen in the gpt-3.5-turbo (with context), has a substantial positive impact on translation quality. This emphasizes the importance of context-awareness in machine translation tasks, as it enables models to generate more contextually coherent and accurate translations.

While the introduction of context, as observed in the gpt-3.5-turbo (with context) model, significantly enhances translation quality, it is crucial to recognize that this performance boost is contingent upon the availability of relevant contextual information. In practical scenarios, contextual information may not always be readily accessible or clearly defined. Machine translation often operates in diverse and dynamic contexts, ranging from general conversation to highly specialized domains. In some instances, such as open-domain conversations or translations of standalone sentences, providing context may be challenging or impractical. In such cases, the model's reliance on context might not yield substantial improvements in translation quality.

Table 2: Translation evaluation on BLEU score

Model	BLEU (mean)
vinai-translate-en2vi	0.2814
gpt-3.5-turbo	0.2951
gpt-3.5-turbo (with context)	0.4049

5 Future Work

We recognize several promising avenues for future research and development that can further enhance the capabilities and applications of RAG pipeline in the context of healthcare and natural language understanding. The following areas represent potential directions for future work:

5.1 Retrieval Performance Improvement

One notable avenue for investigation is the observed phenomenon where our fine-tuned embedding model, vishc-med-qa, performs better in semantic textual similarity, yet seemingly decreases the performance of the retriever component when compared to the base model.

5.2 Enhancing the Generator Component

The generator component within the RAG pipeline plays a pivotal role in producing coherent and contextually relevant responses based on retrieved information. Future work can focus on the following aspects:

Fine-Tuning and Architecture Improvements: Further research can delve into fine-tuning strategies and architectural enhancements for the generator model. This may involve exploring novel pre-training objectives, model architectures, or training techniques to optimize response generation.

Multimodal Capabilities: Investigating the integration of multimodal capabilities, such as combining text and images, can extend the pipeline's utility to tasks that require a richer understanding of both textual and visual information.

5.3 Evaluating Bilingual Medical Embedding Models on Downstream Tasks

Future research can extend the evaluation of this model to various downstream tasks within the medical domain, including:

Named Entity Recognition (NER): Assessing the effectiveness of the embedding model in identifying medical entities, such as diseases, medications, and anatomical structures, can provide insights into its utility for information extraction tasks.

Machine Translation: Expanding the evaluation to machine translation tasks involving medical texts can help determine the model's efficacy in facilitating accurate and contextually relevant translations within the healthcare domain.

6 Conclusion

This research has explored the potential of Retrieval-Augmented Generation (RAG) pipelines in addressing medical question-answering tasks, emphasizing cross-lingual EN-VI information retrieval and question-answering.

In our experiments, we evaluated the performance of various models and methods including embedding with our fine-tuned model using knowledge distillation, retriever component with 2 embedding models and hybrid search and the introduction of contextual information significantly improved translation quality.

References

[1] Kamal raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. BioSimCSE: BioMedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and

Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[3] Nhung Thi-Hong Nguyen, Phuong Phan-Dieu Ha, Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Spbertqa: A two-stage question answering system based on sentence transformers for medical texts, 2022.

[4] Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*, 2022.

[5] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. Kilt: a benchmark for knowledge intensive language tasks, 2021.

[6] Fabio Petroni, Tim Rocktäschel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? *CoRR*, abs/1909.01066, 2019.

[7] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271, 2020.

[8] Nils Reimers, Philip Beyer, and Iryna Gurevych. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[9] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[10] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November 2020. Association for Computational Linguistics.

[11] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *CoRR*, abs/2002.08910, 2020.

[12] Nhu Diep Vo. Vinuni evmed - vinuni cecs internal dataset, 2023.

A Appendix A: Medical Q&A Chatbot

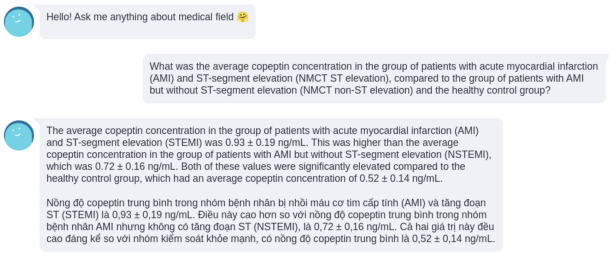

Figure 4: Medical Q&A Chatbot Demo

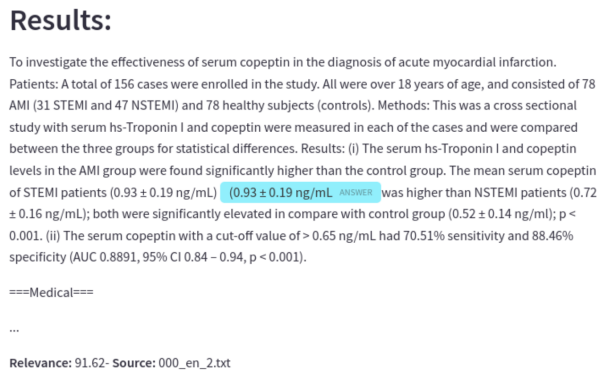

Figure 5: Relevant context for answer generation