

Enhancing RAG Pipeline Performance with Translation-Based Embedding Strategies for Non-English Documents

Can Iscan*, Muhammet Furkan Ozara*, Akhan Akbulut*[†]

*R&D Center of Next4biz Sahrayicedit Mah, Pakpen Plaza, No:40/4, 34734 Istanbul, Turkey
can.iscan@next4biz.com, furkan.ozara@next4biz.com, akhan.akbulut@next4biz.com

[†]Department of Computer Engineering Istanbul Kültür University 34536 Istanbul, Turkey
a.akbulut@iku.edu.tr

Abstract—Despite the advances in multilingual embedding models, they often underperform compared to embedding documents in English, which limits the effectiveness of Retrieval-Augmented Generation (RAG) systems in non-English question-answering contexts. This disparity poses a significant barrier for non-English speakers to access high-quality AI-driven retrieval systems. To address this challenge, this paper proposes a novel approach that leverages English embeddings enhanced by translation. Our approach involves translating texts into English and then embedding them while preserving the original information as metadata. By utilizing robust English and multilingual embedding models, we achieve superior results in RAG systems. We present a comprehensive approach for constructing a question-answering system using GPT-4o technology, which leverages translation models to enhance English embeddings. We conducted actual studies to compare the suggested methodology with conventional methods using various embedding models, employing the RAGAS framework. Initial findings indicate that our method greatly enhances the retrieval effectiveness of RAG systems in non-English environments, attaining higher context precision and recall metrics. This adaptable solution incorporates English embeddings into multilingual applications, providing a versatile and effective method to improve non-English scenarios by using the capabilities of translation models.

Index Terms—RAG, question-answering, multilingual embeddings, translation, ChatGPT, GPT-4o

I. INTRODUCTION

Today's world is getting revolutionized by the advancements made in artificial intelligence, machine learning, generative adversarial networks, and natural language processing [1]. In the past, no one would believe that a mathematical formula inside some chips could compose music, draw glamorous artworks, generate photorealistic video clips, write a hundred pages long novels, solve the hardest math problems, and drive cars on highways [2]. Among all of those capabilities that the current state of machine learning models can do, the most fascinating and may be the most helpful one is chatbots [3].

The popularity of chatbots rised at the end of 2022 with the release of ChatGPT. This product showed people how far artificial intelligence has come by its knowledge of nearly the whole internet, the intelligence of understanding questions and with its replies, also how practical it is [4]. The underlying

technology powering these chatbots is comprised of sophisticated language models, which are trained using vast amounts of textual data sourced from various internet platforms such as blogs, encyclopedias, e-books, and forums.

Chatbot systems are being more commonly included into different platforms including Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and Customer Service Management (CSM) solutions to offer effective and scalable customer support. These solutions are incredibly practical since they aid clients in meeting their needs, enhancing overall satisfaction and efficiency. We employ data from the Next4Biz CSM platform to assess our methodology in this study. The evaluation is carried out using human-generated questions that are exclusively focused on Next4Biz CSM, guaranteeing the pertinence and precision of our assessment.

Large language models (LLMs) and chatbots exhibit high performance when provided with training data that covers a substantial percentage of the internet. Nevertheless, adopting new information into LLMs continues to be a difficult task, despite the utilization of methods such as fine-tuning [5]. For example, a chatbot that is included into a company's customer care platform may lack specialized information regarding the company's products. One potential solution to this issue is the implementation of Retrieval Augmented Generation (RAG) pipelines [6].

RAG pipelines facilitate the acquisition of new knowledge by chatbots in particular areas by finding the most pertinent documents and integrating them into the input for the LLM. This method has enabled the invention of sophisticated question-answering systems, bots for customer service, and tools for content production and summarization. The embedding models are a fundamental element of RAG systems as they are responsible for extracting the semantic meaning of the input text. While embedding models are created for both pure English and multilingual domains, they tend to exhibit superior performance in English, even when the model is designed to handle many languages [7]. Utilizing powerful English embedding models or attaining the same level of performance as English embeddings in multilingual models for other languages can improve the performance of RAG

applications.

This research introduces a new approach to enhance the effectiveness of English embeddings in multilingual models. It also explores the utilization of English embedding models for non-English documents to enhance the performance of RAG in question-answering domains. We utilize multiple English and multilingual embedding models in a question-answering pipeline powered by GPT-4. We evaluate their performance using the RAGAS framework. Our methodology consists of employing translation models to convert texts into English while preserving the original language version as metadata. We then do retrieval using the English-translated versions and utilize the untranslated version within the appropriate context.

The primary contributions of this research are four-fold:

- We demonstrate the potential of English embedding models to improve RAG performance in multilingual scenarios by using translation.
- By translating non-English documents into English and embedding them with multilingual embedding models, we demonstrate the performance advantages over traditional methods.
- We benchmark multiple English and multilingual embedding models using the suggested approach to find the most efficient and accurate RAG pipeline models in non-English question-answering systems and compare their performance.
- We provide a comprehensive analysis of the cost-benefit ratio of implementing the proposed RAG pipeline, offering valuable guidance for practical applications and future research.

The paper is organized as follows. In Section II, we discuss the related research work in the RAG field and highlight similar approaches to achieve better retrieval performance. Section III explains the proposed model and the techniques that are employed. In Section IV, we analyze the results of the experiments, comparing the performance of different embedding models and discussing the impact of using English embeddings on non-English question-answering tasks. Section V presents an in-depth analysis of the findings, and possible constraints and discusses the applicability and scalability of the solution. In Section VI, we summarize the key points of this research and give future research directions.

II. RELATED WORK

Researchers and AI engineers from industry have made tremendous progress in the topic of RAG, researching various use cases to improve the capabilities of LLMs. RAG systems enhance the precision of LLM-generated answers and also facilitate the use of LLMs in new or specific fields without requiring lengthy retraining. RAG systems surpass the constraints of LLMs that are exclusively trained on existing data by including supplementary documents like QA text pairings or complete books. The vast range of applications, such as customer care bots, complex question-answering systems, and content production tools, has resulted in the widespread use of RAG.

Researchers and AI engineers from the industry have investigated many applications of RAG in the literature. In addition to enhancing the precision of responses given by LLMs, RAG allows LLMs to be utilized for novel or specific applications without requiring additional training. RAG systems enhance the capabilities of LLMs by including other documents such as QA text pairs or complete books. This expansion enables them to produce text that is more contextually appropriate and precise, surpassing the limitations of their training data. RAG systems have broad applications in various domains, including customer support bots, sophisticated question-answering systems, and content production tools.

Prior to the introduction of RAG approaches in the literature, major improvements in the field of natural language processing (NLP) and embedding models established the foundation for the advanced RAG systems that are currently employed. Mikolov et al. were the first to develop embedding models, which convert texts into low-dimensional vectors that capture semantic meaning. They introduced the 'word2vec' model as a way to learn word representations [8]. Later developments included the introduction of models such as GloVe [9] and fastText [10]. Although the majority of these models were first trained using English data, subsequent development led to the creation of multilingual embedding models that can function across many languages. Nevertheless, these multilingual models generally exhibit worse performance in comparison to models purely trained in English [11].

Several approaches have been suggested to get better performance from RAG on multilingual documents. The most common way is to use a multilingual embedding model, but these models are generally considered subpar compared to their English-only counterparts [12]. Another study introduced a framework for multilingual keyphrase generation with iterative retriever-generator training, further illustrating the potential of leveraging advanced techniques in multilingual NLP tasks [13].

Translation-based RAG methods involve translating the documents from the target language to English and can utilize the English embedding models. Again, this demands very high-quality translation models as the goodness of the translations affects directly on the quality of the overall performance [14]. Recent studies now suggest that large-scale multilingual NMT models, trained using large amounts of parallel data across a large number of languages, can indeed raise translation quality for both high- and low-resource languages. For instance, Google has recently conducted work on massively multilingual NMT; it found that considerable improvements in quality arise when models are trained on a large, highly multilingual corpus [15].

Researchers have investigated human-like translation strategies using LLMs, which have shown to improve translation quality. Techniques such as Multi-Aspect Prompting and Selection (MAPS) replicate the nuanced processes of human translators, offering more accurate translations by utilizing context and minimizing issues like hallucinations and ambiguities. These advancements highlight the necessity of high-

quality translation models and sophisticated methods for effective RAG systems [16].

In our opinion, RAG has greatly improved LLMs, allowing them to be used in many fields without lengthy retraining. The cornerstone for these advances was early embedding models like 'word2vec,' GloVe, and fastText, which were trained on English data. Mixed-language models generally underperform English-only models. Innovative translation tactics and translation-based methodologies strive to close this gap, showing RAG systems' potential for customer support and content production.

III. METHODOLOGY

To investigate the effectiveness of using English embeddings for non-English documents in RAG pipelines, we designed a comprehensive experimental setup involving real-world documents, a vector store, a translation mechanism, an evaluation framework, and multiple embedding models.

A. Documents

This research involved analyzing 88 Turkish documents from the Next4biz CSM user manual, which contain instructions and descriptions for the software products. The aim of the question-answering system was to deliver precise responses to user questions about the Next4biz CSM software. We developed a RAG pipeline that allows the LLM to access and extract pertinent information from these documents to generate accurate answers.

B. Vector store

Vector stores are one of the key components of the RAG systems. Vector stores, or vector databases, provide an efficient way to store and retrieve high-dimensional vectors. In general, vector stores are used with an embedding model to allow storing and retrieving processes more convenient. These vector embeddings represent text, images, or other data types in a dense format that captures semantic meaning, enabling rapid similarity searches. By leveraging vector stores, RAG systems can efficiently handle large volumes of data, quickly find relevant information, and provide more accurate and relevant responses. In this study, we took advantage of one of the most popular vector stores, ChromaDB [17].

ChromaDB is an open-source vector store used to store and retrieve vector embeddings. It is highly optimized for performance and scalability, making it the optimal choice for RAG applications. Additionally, It can be seamlessly integrated with popular machine learning frameworks and embedding models, providing flexibility in generation and retrieval steps. In our study, leveraging ChromaDB allows us to fairly benchmark the proposed approach with the chosen embedding models.

C. Embedding Models

Embedding models are machine learning models which compress the high-dimension data such as texts and images, into lower-dimension vectors. These produced vectors capture the semantic meaning of the input, which allows to

representation of similar concepts with non-identical inputs. Embedding models are highly used in natural language tasks such as translation, classification, summarization, and relevant text retrieval.

Embedding models work by learning a mapping function that reduces high-dimensional data into a lower-dimensional space. It is intended to understand the meanings of various terms in context which means it tries to understand what different words mean when used together. This process is achieved through examinations of how frequently two words appear with each other within a set of texts known as a corpus. Words that co-occur often tend to have related meanings or synonyms. Using such ideas, the embedding matrix would place these terms close together in Euclidean space.

In the context of question-answering applications, embedding models are utilized to embed documents that contain the necessary information for answering and also embed the question to find the closest documents to the asked question. Afterwards, those pertinent documents are passed to LLM with the question, so that the answer can be correctly generated using the selected documents. By locating an embedding model within a vector store, the documents can be easily embedded and stored, also the most relevant documents can be efficiently retrieved during question-answering.

Embedding models are categorized based on the languages they are designed to process. Multilingual embedding models are trained on big volume of text including various languages which allows them to represent words and concepts across different languages in a single embedding space. However, multilingual models might struggle to capture some of the meaning from the text compared to English embedding models which are trained with corpus containing only English. The choice between these type of embedding models are made depending on the application and language requirements.

In our research, we investigate the performance gain of utilizing English embeddings in a question-answering pipeline for non-English languages such as Turkish, through the application of English and multilingual embedding models. For this purpose, we included several multilingual and English models to the experimental setup. In Table I, the tested embedding models are listed with their benchmark result according to Massive Text Embedding Benchmark (MTEB) [18].

TABLE I
EXPERIMENTED EMBEDDING MODELS

<i>Model Name</i>	<i>Language</i>	<i>Model Size (Million Params)</i>	<i>Embedding Space</i>	<i>Retrieval Score (English)</i>
OpenAI/ text-embedding-3-large	Multilingual	-	3072	55.44
OpenAI/ text-embedding-3-small	Multilingual	-	1536	51.08
Alibaba-NLP/ gte-large-en-v1.5	English	434	1024	57.91
intfloat/ multilingual-e5-large	Multilingual	560	1024	51.43
Snowflake/ snowflake-arctic-embed-l	English	334	1024	55.98

D. Proposed Approach

RAG pipelines used in question-answering systems generally follow similar architecture to each other. A traditional RAG pipeline has 4 main roles which are user, vector store, embedding model and LLM. The flow of RAG pipeline starts with receiving a question from the user. This question is passed to a retrieval system, which is made from a vector store, to run a similarity search and retrieve the most relevant texts to the question. Then, the retrieved paragraphs and the question is passed to LLM to generate an answer. Lastly, the answer generated from a LLM send back to the user as response. This flow is depicted in Figure 1.

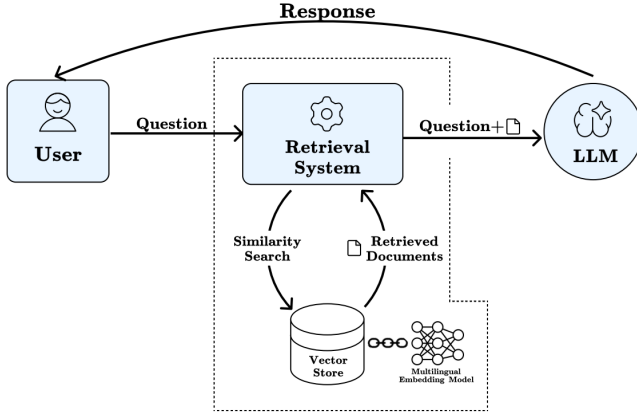


Fig. 1. Traditional RAG pipeline

When the traditional approach is utilized to handle non-English documents, a multilingual embedding model has to be used. The performances of multilingual models are relatively low compared to pure English models. Also, some of the multilingual models achieve better retrieval performances in English compared to non-English languages. These problems arouse a need for embedding non-English documents as they were written in English for both English and multilingual models. We have tackled this problem by locating a translator in front of vector space.

In our proposed approach, the utilization of English embeddings for non-English documents is made allowed by harnessing the power of state-of-the-art translation models. Before letting the pipeline to answer questions, the vector store is required to be populated with the documents which are translated to English with the selected translation model. The question-answering flow of our proposed model starts by retrieving a question from the user and passed to the retrieval system. Then, the question is translated to English with the selected translation model. The similarity search has been done by embedding the translated question and selecting up the closest documents. Afterwards, the original versions of the retrieved documents have been obtained from the metadata attachments. The question and the retrieved documents are packed together to be sent to LLM for answer generation. This flow of the proposed approach is presented

in Figure 2

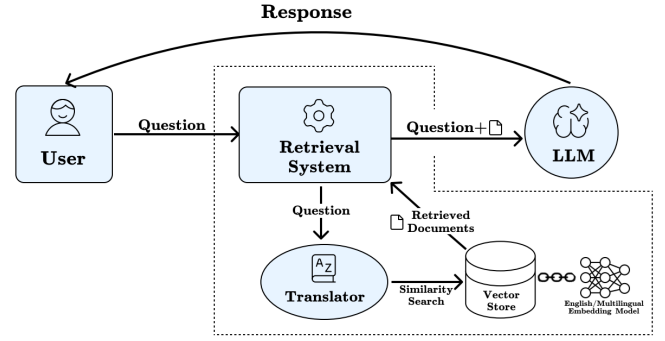


Fig. 2. Translator adapted RAG pipeline

IV. EXPERIMENTAL RESULTS

In this section, we present the results of our experiments. The evaluation of the proposed approach as opposed to traditional RAG approach for non-English use cases have been evaluated with RAG Assessment (RAGAS) framework [19]. RAGAS is a popular benchmarking framework for assessing RAG pipelines. To find out how well our approach is doing, we have integrated RAGAS with our RAG pipeline that we prepared for this research. RAGAS is taking the advantage of LLM models to evaluate the response of question-answering bots by asking questions regarding to retrieved context and the question such as if the retrieved context is sufficient to answer the asked question. We have combined RAGAS with GPT-4o, since we also use that LLM model in response generation. Using a smaller LLM model in RAGAS while using a larger model in RAG pipeline causes misleading metric results. We have mainly focused on two metrics related to retrieval performance explained below.

a) *Context Recall*: Context recall is a measure of how well the retrieved context aligns with the ground truth answer. The metric score falls between 0 and 1 where higher values indicates better performance. To calculate context recall, each sentence in the ground truth answer is analyzed to determine if it can be referenced from the retrieved context. In an ideal scenario, all sentences in the ground truth answer should be covered by the retrieved context.

b) *Context Precision*: Context precision measures whether all of the ground-truth relevant items present in the contexts are ranked higher. It is computed using the question, ground truth, and the retrieved contexts. The metric also ranges between 0 and 1, where higher scores indicate better precision. In an ideal scenario, all relevant chunks should appear at the top ranks of the retrieved contexts.

The experimental process begins with the collection of Turkish documents which is obtained and parsed from the Next4biz CSM products manual. These documents are then translated into English using state-of-the-art neural machine translation model, *Helsinki-NLP/opus-mt-tr-en* from

OPUS-MT project [14]. The translated documents are then processed by English embedding models to generate semantic embeddings. Chroma vector database is employed to store the embeddings of the documents. The original Turkish version of the documents are retained as metadata to be included in the final retrieval step.

The proposed approach and the traditional approach for retrieving non-English documents have been evaluated by utilizing different embedding models to minimize the bias occurring from a embedding model. Multilingual embedding models are tested with both approaches while English models are tested with proposed approach only. 15 different human-written question and related answers are prepared and used with the RAGAS to obtain a measurable result. RAGAS used the RAG pipeline to generate answers for the questions and leveraged GPT-4o to evaluate if the return context is adequate to answer the question.

The findings of the experiments on the proposed approach have shown that the embedding models obtained the same context recall score of 0.76. The best performing model among selected ones is happened to be OpenAI’s text-embedding-large-3 with the context recall score of 0.76 and context precision of 0.65. Among the embedding models tested, the performance difference is happened on the ordering of the returned documents which is represented with context precision metric. In Table II, the experimental results of the traditional approach is shown.

TABLE II
EVALUATION RESULTS OF TRADITIONAL APPROACH

<i>Model Name</i>	<i>Language</i>	<i>Context Recall</i>	<i>Context Precision</i>
intfloat/ multilingual-e5-large	Multilingual	0.76	0.64
OpenAI/ text-embedding-3-large	Multilingual	0.76	0.65
OpenAI/ text-embedding-3-small	Multilingual	0.76	0.61

According to the evaluation results of the proposed approach, OpenAI’s text-embedding-3-large model acquired the best metric scores of 0.81 in context recall and 0.72 in context precision. The second best-performing model is found to be *gte-large-en-v1.5* which closely follows the champion model in metric scores also much smaller in embedding dimension size. The embedding model from Snowflake obtained the lowest metric scores, but it is also the smallest model in parameter size among the selected models. The results of the proposed approach are presented in Table III.

When both approaches are compared with the results of the assessment, OpenAI’s *text-embedding-3-large* model stands out with the best context recall and precision scores of 0.81 and 0.72. The proposed approach brought a %10 performance gain compared to the traditional approach in the text-embedding-3-large model. However, when intfloat’s *multilingual-e3-large* model is employed within the proposed

TABLE III
EVALUATION RESULTS OF PROPOSED APPROACH

<i>Model Name</i>	<i>Language</i>	<i>Context Recall</i>	<i>Context Precision</i>
intfloat/ multilingual-e5-large	Multilingual	0.76	0.58
Alibaba-NLP/ gte-large-en-v1.5	English	0.80	0.69
Snowflake/ snowflake-arctic-embed-l	English	0.44	0.28
OpenAI/ text-embedding-3-large	Multilingual	0.81	0.72
OpenAI/ text-embedding-3-small	Multilingual	0.76	0.71

approach, a drop in context precision is observed. The English embedding model *gte-large-en-v1.5* achieved better metric scores compared to multilingual-e5-large, even though it has fewer parameters than the other. Also, it achieved a better context recall score than the text-embedding-3-small model.

V. DISCUSSION

In the current state of question-answering pipelines, Retrieval Augmented Generation is one of the most used techniques to feed LLM models with new information. The most crucial part of the RAG pipelines is embedding models used while storing and retrieving the relevant documents. Employing larger and more advanced embedding models improves the retrieval of more similar documents to the question, leads to more accurate responses. Within the experiments conducted in this research work, several different embedding models have been evaluated with the proposed and traditional RAG approaches to find out if the translation of non-English documents improves the retrieval performance. The assessment presents that the multilingual embedding models, such as OpenAI’s text-embedding-3-large, do not perform as well as with Turkish documents compared to English documents. Also, the smaller English embedding model accomplished better results compared to other multilingual models with bigger parameter sizes. The findings have revealed that the translation of non-English documents to English has a positive impact on retrieval performance when utilizing either a multilingual or English embedding model.

The results of the experiments show that multilingual models exhibit greater performance in the English language compared to other or less-used languages like Turkish. The retrieval scores obtained in the experiments have revealed that the information loss occurring in the translation of the documents is not higher than the performance gain acquired using the proposed approach. The utilization of better translation models/techniques could decrease the information loss, allowing the proposed approach to perform even better in retrieval. The importance of embedding model selection is underscored by the experiments carried out. The experiments revealed significant performance differences among English

and multilingual embedding models themselves. This highlights the need for a careful selection process tailored to the specific requirements of the application. Concerning this study, models having less number of parameters yet optimized well for particular tasks proved more effective than bigger generalized ones. For professionals who would want to apply similar techniques, an understanding of this kind could be beneficial.

The experiments in this study are constructed with Turkish documents which obtained promising results when the proposed approach was applied. The results and the best-performing models can vary when the approach is tested with other languages. However, the proposed approach is seen to perform better than the traditional approach for less-used languages. The data used for the training of the embedding model is the most determinative factor in this case. If the data contains the used language in a great volume, the performance difference between it and English tends to be smaller. The response times between traditional and proposed approaches vary slightly due to locating a translation model in front of the vector store. The translation model needs to translate the incoming questions into English in real-time. The retrieved documents from the vector store do not need to be translated since their original versions are stored in metadata. The response time difference between approaches occurs only from the translation of the questions, which are generally a sentence long. Thus, translation does not bring a speed concern unless much bigger models are utilized for translation.

VI. CONCLUSION

In this work, we developed an innovative strategy to enhance RAG performance by shifting retrieval processes to English for non-English documents through the application of translation models. We evaluated this strategy using multiple embedding models and the RAGAS framework, testing it with 15 question-answer pairs. Our findings show that this approach significantly outperforms traditional methods in terms of context precision and recall. The OpenAI text-embedding-3-large multilingual model achieved a context recall of 0.81 and a context precision of 0.73 with our method, marking a 10% improvement over the traditional method's context recall of 0.76 and context precision of 0.65. This advancement underscores the effectiveness of leveraging high-quality English embedding models for non-English tasks and suggests that further enhancements could be made with more advanced translation models.

In future studies, the applicability of this approach can be tested for other languages than Turkish. Additionally, the proposed approach can be assessed further with utilizing different translation models to investigate the impact of translation models.

REFERENCES

- [1] Immad A Shah and SukhDev Mishra. Artificial intelligence in advancing occupational health and safety: an encapsulation of developments. *Journal of Occupational Health*, 66(1):uiad017, 01 2024.
- [2] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.
- [3] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentaoh Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401, 2020.
- [4] Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. Chatgpt: A meta-analysis after 2.5 months, 2023.
- [5] Andy Wang, Cong Liu, Jingye Yang, and Chunhua Weng. Fine-tuning large language models for rare disease concept normalization. *Journal of the American Medical Informatics Association*, page ocae133, 06 2024.
- [6] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *CoRR*, abs/2202.01110, 2022.
- [7] Chihiro Yano, Akihiko Fukuchi, Shoko Fukasawa, Hideyuki Tachibana, and Yotaro Watanabe. Multilingual sentence-t5: Scalable sentence encoders for multilingual applications, 2024.
- [8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, January 2013.
- [9] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- [11] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.
- [12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.
- [13] Yifan Gao, Qingyu Yin, Zheng Li, Rui Meng, Tong Zhao, Bing Yin, Irwin King, and Michael R. Lyu. Retrieval-augmented multilingual keyphrase generation with retriever-generator iterative training, 2022.
- [14] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November 2020. European Association for Machine Translation.
- [15] Sneha Reddy Kudugunta, Ankur Bapna, Isaac Caswell, Naveen Arivazhagan, and Orhan Firat. Investigating multilingual NMT representations at scale. *CoRR*, abs/1909.02197, 2019.
- [16] Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246, 03 2024.
- [17] Chroma Team. Chroma documentation. Online, 2024. docs.trychroma.com.
- [18] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023.
- [19] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.