ELABORATING DRUG DATA INFORMATION RETRIEVAL EFFICIENCY USING OPTIMIZED CLUSTERING TECHNIQUES

Maryamah Maryamaha, Muhammad Maula Irfanib, Netri Alia Rahmic

a,b,c Data Science Technology, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya, Indonesia

**Abstract**

**Aim**:

In the age of digital transformation, educational institutions are increasingly embracing technology to improve accessibility and communication. Within this context, our faculty recognizes the imperative to address the diverse needs of users navigating our online platform. To meet this challenge, we have developed a versatile chatbot tailored to not only assist prospective students but also to engage a wider audience interested on exploring both our academic and non-academic informations. This paper delves into the developing of this comprehensive chatbot, with the goal of providing a seamless and informative experience for all website visitors.

**Method**:

The development of our chatbot employs the Retrieval-Augmented Generation (RAG) technique which combining pre-trained parametric and non-parametric memory for response generation (Lewis et al., 2020). RAG comprises a Retrieval module and an LLM-based generation module (Es et al., 2023). The Retrieval module systematically identifies documents stored in the vector database that are most likely to contain relevant information related to the user's query. The selected documents, along with the user's query, are then presented to the LLM as part of the prompt. This enables the model to generate responses based on the knowledge provided in the relevant documents.

To determine the most effective retrieval method, we compared similarity search and MMR Search (Carbonell, 1998). The distinction lies in the fact that while similarity search focuses solely on similarity, MMR Search also emphasizes document diversity, thereby reducing redundancy in the selected documents. Both retrieval methods were then applied to three embedding techniques: OpenAI Ada (OpenAI, 2022), Google Embedding-001, and MPNet-Multilingual (Reimers & Gurevych, 2020), considering our dataset is in the Indonesian language. Knowledge about our faculty was obtained from faculty documents and the website, subsequently stored in the ChromaDB vector database. The best retrieval method was utilized to extract context for answering user queries inputted into the LLM. We compared two LLMs, namely OpenAI GPT-3.5-Turbo and the latest Google Gemini Pro (Google Deepmind, 2023).

**Result and Discussion:**

From the combination of these experiments, we found that using similarity search on the vector representation of OpenAI Ada demonstrates its ability to understand the semantic meaning in text better than other methods. This result is based on a dataset labeled by our students measured with Recall and Precision metrics. We also discovered that the use of MPNet-Multilingual is a viable alternative due to its efficiency in terms of time and cost compared to the paid OpenAI Ada. On the LLM-based generation side, we found that OpenAI GPT-3.5-Turbo outperformed Google Gemini in response, measured by BLEU Score (Papineni et al., 2001b) and ROUGE Score (Lin, 2004). We identified potential for improvement in this chatbot by enhancing the knowledge provided to the LLM in future iterations.

**Conclusion:**

In conclusion, our developed chatbot demonstrates significant potential to revolutionize the information processes within the faculty, offering a seamless experience for users. The successful implementation of the Retrieval-Augmented Generation technique, optimal retrieval method selection, and the superiority of OpenAI GPT-3.5-Turbo in response generation underscore the chatbot's capability to enhance accessibility and communication. The cost-efficiency demonstrated by MPNet-Multilingual indicates that the chatbot is also an economically viable option.

Keywords: Chatbot, Retrieval-Augmented Generation, Google Gemini, Large Language Model, Information Retrieval

References:

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). Retrieved from <https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf>

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.2309.15217>

Google Deepmind. (2023). Gemini - Google DeepMind. Deepmind.google. <https://deepmind.google/technologies/gemini>

OpenAI. (2022, December). New and improved embedding model. Openai.com. <https://openai.com/blog/new-and-improved-embedding-model>

Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. ArXiv:2004.09813 [Cs]. <https://arxiv.org/abs/2004.09813>

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001b). BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02. https://doi.org/10.3115/1073083.1073135

Lin, C.-Y. (2004, July 1). ROUGE: A Package for Automatic Evaluation of Summaries. ACLWeb; Association for Computational Linguistics. https://aclanthology.org/W04-1013Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001b). BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02. https://doi.org/10.3115/1073083.1073135