International Research Journal of Education and Technology

Peer Reviewed Journal

ISSN 2581-7795

RETRIEVAL AUGMENTED GENERATION FOR TAMIL

Gokul S

[1]Studunt, Dept. of Artificial Intelligence and Machine Learning, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India

---

Abstract - *Recent advances in natural language processing (NLP), especially in retrieval augmented generation (RAG), have greatly increased response accuracy through integrated information retrieval and information processing but so this technology mainly focuses on widely spoken languages like English, leaving languages like Tamil less available. This study aims to improve language technology for Tamil speakers by developing a specialized RAG model for Tamil. This model was developed to address the unique grammatical and cultural aspects of Tamil. By training the Tamil text, the model can give more accurate and relevant answers. It performed well in simple queries but had trouble with words with multiple meanings, regional accents and old Tamil. Even if possible, more changes are needed to address complex questions. Finally, this study shows that the Tamil-specific RAG model can make language technology more favorable for Tamil speakers, so that the responses are authentic and culturally appropriate and further developments enable the model to work well.*

Key Words: *Tamil Language, Retrieval Augmented Generation (RAG), Natural Language Processing (NLP), Language Model, Information Retrieval, Text Generation, Machine Learning.*

1.INTRODUCTION

The rapidly growing rise of natural language processing (NLP) has changed the way machine and human language interact, enabling smarter and more accurate communication Despite tremendous progress, most NLP technologies are designed to consume languages as widely spoken as English, and languages like Tamil have been left in short supply. Tamil language processing poses unique challenges to traditional NLP models due to its complex grammar structure, cultural complexity, and local variability, which often complicates standard language techniques These factors provide it is difficult to handle Tamil properly for traditional models trained mainly in simple structured language. This study focuses on a Tamil-specific NLP model using Retrieval Augmented Generation (RAG) techniques, an approach that combines information retrieval and text generation to improve response accuracy and relevance. This technology can better serve the Tamil-speaking communities, improving services such as customer support, education and information retrieval.

1.1 Background of the Work

The development of a Tamil-specific RAG-based NLP model aims to address the unique challenges posed by the Tamil language in natural language processing. Tamil, with its intricate grammar, rich cultural heritage, and diverse regional dialects, presents significant difficulties for standard NLP systems primarily designed for widely spoken languages. Traditional language models often fail to capture the linguistic nuances and cultural context essential for accurate communication. This project seeks to bridge this gap by integrating Retrieval Augmented Generation (RAG) techniques, which combine information retrieval and text generation to enhance the relevance and accuracy of responses. Key components of the project include adapting RAG methods to suit Tamil's specific grammatical structure, incorporating culturally sensitive context, and ensuring the model's ability to handle regional variations and classical forms of Tamil. The project will also involve the deployment of large Tamil datasets for model training and evaluation, aiming to improve Tamil language technology for a wide range of applications.

1.2 Motivation and Scope of the Proposed Work

The development of the Retrieval Augmented Generation (RAG) model for Tamil aims at creating customized Natural Language Processing (NLP) solutions that address the specific linguistic needs and cultural context of Tamil speakers. Tamil, with its complex grammar, regional variations, and rich cultural heritage, presents unique challenges for traditional NLP systems. This project seeks to modify RAG specifically for Tamil to facilitate accurate, context-aware responses, making it more effective for various applications like automated customer support, education, and information retrieval. By adapting RAG methods to Tamil, the goal is to support digital inclusion, allowing Tamil-speaking communities to access advanced NLP technology that is simple, reliable, and culturally sensitive. The proposed model will integrate information retrieval and text generation to improve response accuracy, ensuring that it meets the linguistic and cultural nuances of Tamil. The project will involve training the model with large Tamil datasets, refining its parameters, and overcoming challenges like regional

ISSN 2581-7795

dialects and classical Tamil usage. Expert guidance from NLP specialists will help optimize the model for real-time performance, contributing to its continuous improvement for diverse Tamil language applications. Success will be determined by the model's ability to deliver precise, culturally relevant, and contextually appropriate responses, thereby enabling more inclusive language technology for Tamil-speaking communities.

2. METHODOLOGY

The methodology for implementing the Tamil RAG system is divided into two primary phases such as preparation and processing. Each phase plays a critical role in ensuring the system performs efficiently and delivers high-quality responses.

2.1 Preparation Phase

The preparation phase focuses on building a comprehensive Tamil language database for retrieval and generation processes. Key steps include:

Data Collection: Tamil text is gathered from diverse sources like web scraping, PDFs, and publicly available resources.

Preprocessing: Raw text undergoes normalization (standardizing spelling, punctuation), tokenization (breaking text into words or subwords), and vectorization using models like BERT tailored for Tamil.

Embedding Vector Database: The vectorized representations are stored in a database, enabling efficient retrieval of relevant content during user queries.

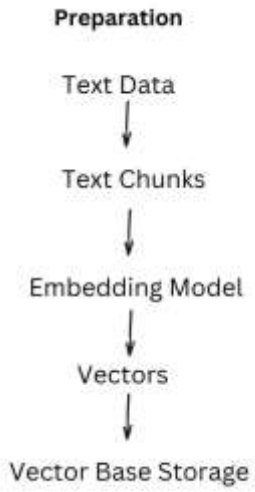

Fig -1- Preparation Phase FLow chart

2.2 Usage Phase

In the retrieval phase, the system fetches relevant text based on the user's query. The steps are:

Query Input and Preprocessing: The user's query is preprocessed (normalized, tokenized, and sometimes translated) to match the format of the stored documents.

Query Vectorization: The query is converted into a vector using the same model applied to the stored documents. This vector represents the query's meaning.

Similarity Search: The query vector is compared to stored vectors using similarity measures, such as cosine similarity. The system finds the closest matching documents.

Retrieving Top-K Results: The system retrieves the top-K most relevant documents based on their similarity to the query.

Post-Retrieval Processing: The retrieved documents are refined to ensure they contain only the most relevant information.

Contextual Answer Generation: A fine-tuned model uses the retrieved information to generate a coherent and accurate response to the user's query.

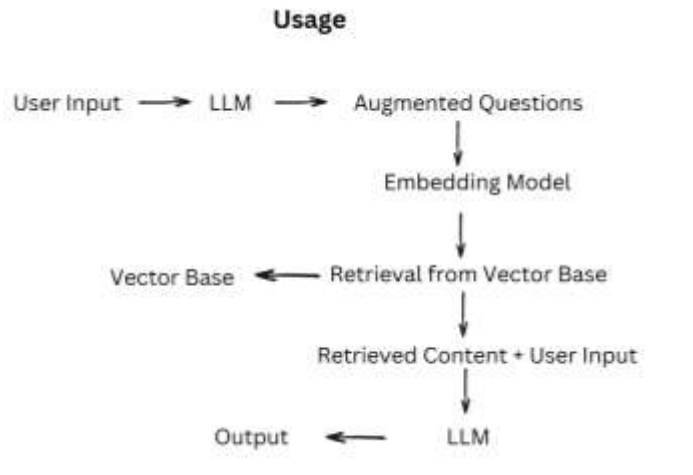

Fig -2- Usage Phase FLow chart

3. CONCLUSIONS

The success of the Tamil-specific RAG model highlights the potential of an adapted NLP system to enhance language technology for Tamil speakers. By addressing the unique linguistic and cultural nuances

ISSN 2581-7795

of Tamil, this model provides an important step forward in improving response accuracy and contextualization. It demonstrates the potential of integrating information retrieval and speech generation, providing Tamil-speaking communities with comprehensive and culturally appropriate digital tools. Ultimately, this model paves the way for further developments in Tamil NLP, providing a more inclusive and effective language technology for a wide range of applications

Suggestions for Future Work

1. **Expanding Data Diversity**: Training the Tamil-specific RAG model on a wider variety of Tamil dialects, regional variations, and cultural contexts can enhance its adaptability and effectiveness across different Tamil-speaking communities..
2. **Integrating Multilingual Support**: Incorporating support for multiple languages and transliterations can make the model more versatile, allowing it to handle queries in different language forms like Tamil-English code-switching or other regional variants.
3. **Improving Handling of Ambiguity**: Enhancing the model's ability to disambiguate words with multiple meanings and better handle context-specific queries would improve its accuracy, especially for complex Tamil expressions.
4. **User Feedback Integration**: Incorporating real-time user feedback for continuous learning and model improvement can help the system adapt and evolve to meet the dynamic needs of Tamil-speaking users.

REFERENCES

[1] Z. Jiang., et al. (2023). Active Retrieval Augmented Generation. *arXiv preprint*, arXiv:2305.06983. arXiv : https://arxiv.org/abs/2305.06983

[2] Salemi, A., & Zamani, H. (2024). Evaluating Retrieval Quality in Retrieval-Augmented Generation. *SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference*,https://dl.acm.org/doi/abs/10.1145/3626772.3657957