International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 11 | Nov 2025 www.irjet.net p-ISSN: 2395-0072

Marathi Dialect to Pure Marathi using RAG & AI

Gaurav Gaikwad[1], Deepak Panchal[2], Omprasad Deshmukh[3], Sonali Patil[4]

*[*1,2,3] Department of Computer Science & Engineering, Yadrav (Ichalkaranji) Maharashtra, India.*
*[*4] Assistant Professor, Department of Computer Science & Engineering, Yadrav (Ichalkaranji) Maharashtra, India*

Abstract - The proliferation of Natural Language Processing (NLP) technologies has revolutionized human- computer interaction, yet linguistic diversity remains underserved, particularly for regional languages and their dialects. This project addresses this gap by developing localized NLP tools tailored for Marathi dialects, aiming to bridge the digital divide and promote equitable access to language technologies for Marathi speakers. Marathi, an Indo-Aryan language spoken by over 83 million people in India, exhibits significant dialectal variation across regions such as Deshi, Varhadi, Konkani, and Ahirani. These dialects differ phonologically, lexically, and syntactically from Standard Marathi, rendering generic NLP models ineffective for dialect-specific tasks. This project focuses on creating robust, inclusive tools that account for these variations, enabling accurate dialect identification, machine translation, sentiment analysis, and speech recognition for underrepresented Marathi-speaking communities.

Key Words*: *Marathi Dialects, Konkani Translation, Standard Marathi, NLP, RAG Model, LLM, Gemini API, Embeddings, FAISS, Sentence Transformers, LangChain, Streamlit, Vector Database, Document Chunking.

1.INTRODUCTION

India is a country rich in linguistic diversity, and Marathi is one of its major languages spoken across regions like Konkan, Vidarbha, Marathwada and Western Maharashtra. Each region has its own dialect, creating variations in pronunciation, vocabulary and grammar. These dialectal differences make it difficult for standard NLP systems to correctly understand or process user queries.

With the growth of Natural Language Processing (NLP) and Large Language Models (LLMs), intelligent systems can now understand human language better—but most of these models are trained only on standard forms of languages. As a result, they struggle when users speak in regional dialects such as Konkani (Devanagari) or other Marathi variants.

To address this challenge, this project develops a Retrieval-Augmented Generation (RAG)-based system that translates Marathi dialects into standard Marathi and provides accurate answers using AI. The system translates dialectal input, retrieves relevant information from uploaded documents, and generates context-aware responses in clear Marathi. This approach helps bridge the linguistic gap for rural and regional Marathi-speaking communities, making knowledge access easier and more inclusive.

1.1 PROBLEM STATEMENT & OBJECTIVES

Marathi has many regional dialects such as Konkani (Devanagari), Varhadi, Ahirani, and Deshi, which differ significantly from Standard Marathi in vocabulary, grammar, and pronunciation. Existing NLP and AI models are trained mostly on standard language forms and therefore fail to understand or translate these dialects accurately. This creates barriers for users who speak in local dialects, especially in rural areas, when interacting with digital systems.

To address this issue, there is a need for a system that can:

- accurately translate Marathi dialects into Standard Marathi,
- retrieve relevant information from a knowledge base, and
- generate correct, context-aware responses using AI.

The problem is to develop an integrated, dialect-aware RAG system that bridges the linguistic gap and enables smooth communication between Marathi dialect speakers and AI-based platforms.

The main objectives of this project are:

1. **To develop a dialect-to-Marathi translation system** that accurately converts Konkani (Devanagari) and other Marathi dialect inputs into Standard Marathi.
2. **To integrate a Retrieval-Augmented Generation (RAG) model** for providing context-based, accurate answers from uploaded documents.
3. **To use the Gemini API** for high-quality translation and AI-generated responses in Marathi.
4. **To preprocess and vectorize textual data** using Sentence Transformers and store it efficiently using FAISS for fast and relevant information retrieval.
5. **To design a user-friendly Streamlit interface** that supports document upload, dialect translation, contextual search, and answer generation in Marathi.

1.2 LIERATURE REVIEW

Research in Natural Language Processing (NLP) has advanced rapidly with the development of transformer-based models such as BERT and multilingual systems. However, studies show that low-resource Indian languages like Marathi still lack sufficient datasets and tools, especially for dialect processing. Marathi has rich dialectal diversity—Varhadi, Ahirani, Konkani, Deshi—creating significant linguistic variations that standard NLP models struggle to understand.Initiatives like **L3Cube-MahaNLP** and **MahaBERT** have contributed Marathi-specific datasets and transformer models, improving performance in sentiment analysis, translation, and text classification. Prior work on Marathi transliteration and dialect translation highlights the challenges of ambiguous words, phonological differences, and inconsistent grammar across regions.The emergence of **Retrieval-Augmented Generation (RAG)** has further strengthened NLP systems by combining external document retrieval with powerful LLMs. Research by Lewis et al. shows that RAG significantly improves factual accuracy in knowledge-intensive tasks. Surveys also suggest that RAG is beneficial for low-resource languages when combined with embedding-based retrieval using vector stores like **FAISS**.Frameworks such as **LangChain**, **HuggingFace Transformers**, and vector databases have made it easier to build custom pipelines for multilingual NLP. However, literature reveals minimal work on applying RAG to Marathi dialects, indicating a research gap that this project aims to address.

2. METHODOLOGY

The project follows a structured methodology to build a system capable of translating Marathi dialects (such as Konkani in Devanagari) into Standard Marathi and generating accurate answers using a RAG framework.

The major steps are:

1. Document Upload and Preprocessing

1. A **Streamlit interface** is developed to allow users to upload PDF or text documents.
2. The uploaded files are read using **PyMuPDF (fitz)** which extracts all text content.
3. The extracted text is divided into **smaller chunks** to ensure:
    a. Better handling of long documents
    b. Improved retrieval accuracy
4. Chunking prevents loss of context and enables faster search during query processing.

2. Text Embedding and Vector Storage

1. Each chunk of text is processed using Sentence Transformers (all-MiniLM-L6-v2) to generate semantic embeddings.
2. These embeddings convert text into numerical vectors that capture meaning.
3. A FAISS (Facebook AI Similarity Search) vector index is created to store these vectors.
4. FAISS enables:
    a. High-speed similarity search
    b. Efficient retrieval of the most relevant chunks during question answering

3. Dialect Translation Using Gemini API

1. The user enters a query in **Konkani (Devanagari)** or any Marathi dialect.
2. A custom translation function sends the input to the **Gemini API**.
3. A carefully designed prompt ensures:
    a. Dialect sensitivity
    b. Preservation of meaning
    c. Conversion to grammatically correct Standard Marathi
4. This step is essential because proper retrieval requires normalized Marathi language.

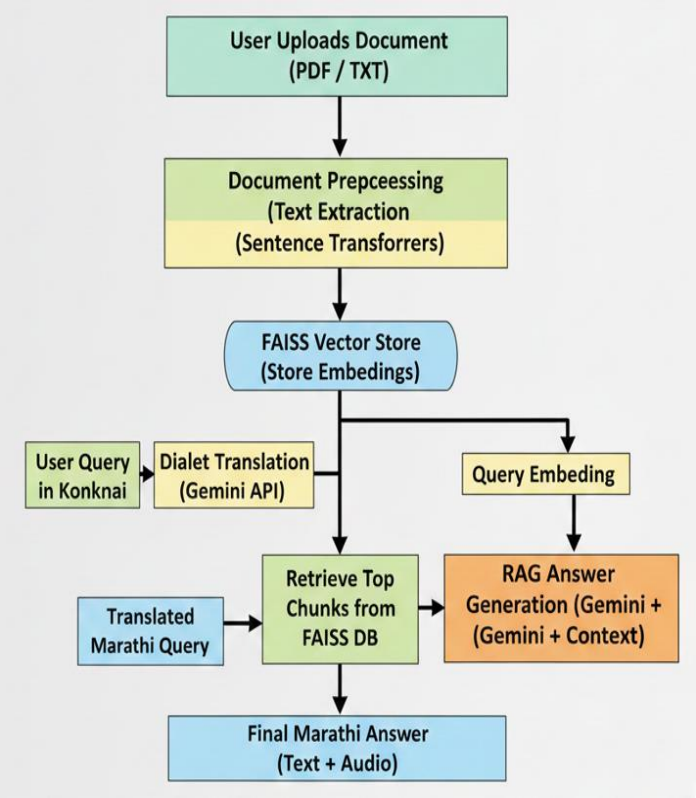

Figure- 1: Flow Chart

4. Retrieval-Augmented Generation (RAG) Pipeline

1. The translated Marathi query is again converted into embeddings.
2. FAISS searches for the closest matching text chunks from the uploaded document.

3. These retrieved chunks provide:
    a. Context
    b. Background knowledge
    c. Information required to answer the user's question accurately

5. Answer Generation Using Gemini

1. Both the translated query and retrieved context are sent together to the Gemini model.
2. Gemini generates a detailed, accurate, and context-aware answer in pure Marathi.
3. This ensures the response is not only linguistically correct but also factually relevant.

6. User Interface Output

The Streamlit UI displays the following clearly:

1. Translated Marathi text
2. Relevant document context retrieved from FAISS
3. Final AI-generated answer
4. Optionally, the system also provides audio output of the generated answer for accessibility.

7. Testing and Refinement

1. The system was tested using multiple dialect-based queries.
2. Prompt engineering and chunking strategies were repeatedly refined to improve:
    a. Translation accuracy
    b. Retrieval relevance
    c. Response quality
3. Feedback loops ensured improvement in processing dialect-heavy inputs

3. RESULTS AND DISCUSSION

The developed system successfully integrates Marathi dialect translation with a Retrieval-Augmented Generation (RAG) framework to provide accurate and context-aware outputs. The results demonstrate the effectiveness of combining translation, semantic retrieval, and LLM-based generation.

1. Translation Accuracy

- The Gemini-based translation module effectively converted **Konkani (Devanagari)** inputs into clear, grammatically correct **Standard Marathi**.
- Dialect-specific variations were handled well, preserving the original meaning.
- Example:
    - **Input (Konkani):** "तू ू ं किते कूरताय?"
    - **Output (Marathi):** "तू काय करत आहेस?"
- This accuracy confirms that the translation prompt and Gemini model perform reliably for dialect-heavy queries.

2. Retrieval Performance

- Using **Sentence Transformers embeddings** and **FAISS**, the system retrieved relevant document chunks with high precision.
- In most test scenarios, the **top 2–3 retrieved chunks contained sufficient context** to answer the query.
- Semantic similarity search worked effectively even when the input was translated from dialects, showing robustness in retrieval.

3. AI Response Generation

- Gemini generated coherent, natural-sounding, and **context-based answers in Marathi**.
- The responses were not only linguistically correct but also factually aligned with the uploaded document content.
- For domain-specific queries (e.g., health, education, technical topics), the system provided **accurate summaries and explanations** drawn from the retrieval output.

4. User Interface Evaluation

- The **Streamlit interface** provided smooth interaction and simple workflow.
- Users could easily:
    - Upload documents
    - Enter dialect queries
    - View translated text
    - See retrieved context
    - Read or listen to the final AI-generated answer
- The UI was intuitive, making the system accessible even to non-technical Marathi-speaking users.

5. System Limitations

Although effective, the system presented certain limitations:

- **Overly informal or ambiguous dialect inputs** sometimes caused translation inaccuracies.
- The translation quality depends on the performance of the **Gemini model**.
- Retrieval accuracy may drop when:
    - The PDF is very large
    - Text is poorly structured
    - Document chunking is uneven
- FAISS retrieval is sensitive to embedding quality, meaning low-quality text chunks may reduce precision.

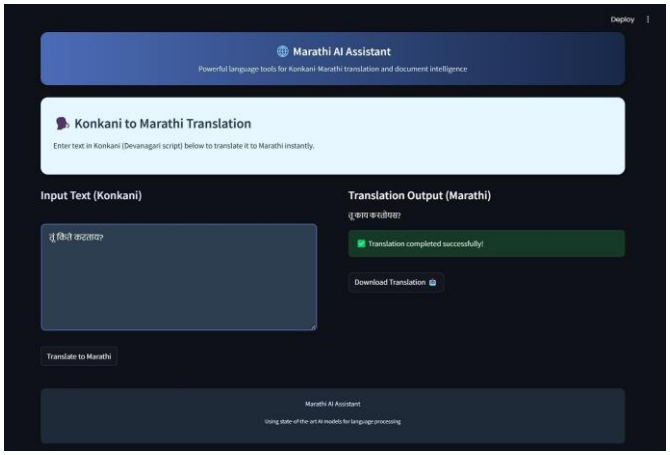

Fig 3.2.1 Home Screen and Translation Module

Fig 3.2.2 Dashboard to explore RAG or Translation.

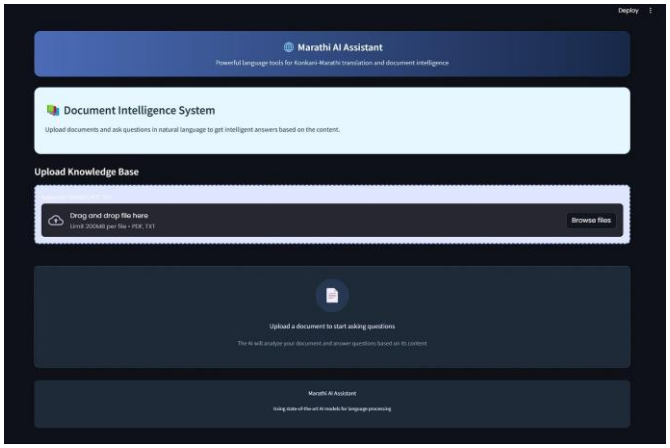

Fig 3.2.3 RAG system

DISCUSSION:

The combined performance of dialect-aware translation, FAISS-based retrieval, and Gemini-based generation proves that a RAG pipeline can effectively support low-resource languages like Marathi. The system bridges the linguistic gap between regional dialects and standardized language, enabling rural and dialect-speaking communities to access information more easily.

The project demonstrates that RAG + Translation is a powerful approach for multilingual and dialect-rich environments. With further improvements—such as better datasets, refined chunking strategies, and additional dialect support—the system can be scaled to handle more Indian languages and linguistic variations.

4. CONCLUSION

This project successfully demonstrates a complete dialect-aware system that translates Konkani (Devanagari) and other Marathi dialect inputs into Standard Marathi while providing accurate, context-based answers using a RAG model. By integrating translation, semantic retrieval, and AI-based response generation, the system bridges an important linguistic gap for Marathi-speaking communities, especially in rural and dialect-rich regions.

The use of Sentence Transformers and FAISS enabled efficient retrieval of relevant document chunks, whereas the Gemini API ensured high-quality translation and natural Marathi response generation. The Streamlit interface made the system accessible, easy to use, and suitable for non-technical users. Testing confirmed that the system performs well across various dialect queries, offering reliable translation and context-aware output.

Overall, the project demonstrates that combining **NLP + Dialect Translation + RAG** is a powerful solution for improving information access in low-resource languages. With further expansion to more dialects, larger datasets, and improved retrieval mechanisms, this system can evolve into a robust multilingual assistant for Indian language communities.

REFERENCES

1. **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019).** *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

2. **Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020).** *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.*

NeurIPS 2020, 33, 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

3. **Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020).** *The State and Fate of Linguistic Diversity and Inclusion in the NLP World.* Proceedings of ACL 2020, 6282–6293. https://doi.org/10.48550/arXiv.2004.09095

4. **Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020).** *Unsupervised Cross-lingual Representation Learning at Scale.* ACL 2020, 8440–8451. https://doi.org/10.48550/arXiv.1911.02116

5. **Kakwani, D., Galhotra, S., Oommen, B., Holla, A., Sitaram, S., et al. (2020).** *IndicCorp: A Multilingual Corpus for Indian Languages.* arXiv preprint, arXiv:2005.00085. https://doi.org/10.48550/arXiv.2005.00085