

CrossLinguaRAG: An Intelligent Cross-Lingual Retrieval System using Multilingual Sentence Embeddings

Guide: Prof. Kiran B
Department of Computer Science and Engineering
Mysuru Royal Institute of Technology (MRIT)
Mandya, Karnataka, India

Sandesh, Yashwanth L
Yashwanth M, Yashwanth P
Department of Computer Science and Engineering
Mysuru Royal Institute of Technology (MRIT)
Mandya, Karnataka, India

Abstract - Digital information in regional languages is expanding rapidly, but accessing such material becomes difficult when the user's query and the available documents are written in different languages. CrossLinguaRAG is a web-based Cross-Language Information Retrieval (CLIR) system developed to retrieve information from Kannada documents using either English or Kannada queries. The system accepts Kannada PDF, DOCX and TXT files and processes them through text extraction, language validation, cleaning and overlapping chunking. Each passage is represented with the pretrained paraphrase-multilingual-MiniLM-L12-v2 Sentence-BERT model. Retrieval uses a hybrid relevance strategy that combines semantic similarity, TF-IDF lexical similarity and keyword matching. The highest-ranked Kannada passages are returned as evidence and can be translated into English. A Retrieval-Augmented Generation (RAG) stage then uses the retrieved evidence and the original query to produce an English response while keeping the supporting passages visible. A manually annotated evaluation set of 20 English and Kannada queries was measured using Recall@1, Recall@3, Mean Reciprocal Rank (MRR) and nDCG@3. The prototype obtained 80.0% overall Recall@1, 90.0% Recall@3, 0.8333 MRR and 0.8695 nDCG@3.

Keywords - Cross-language information retrieval, hybrid information retrieval, Kannada NLP, multilingual Sentence-BERT, retrieval-augmented generation, semantic retrieval.

I. INTRODUCTION

The increasing digitization of regional-language material has created a need for information retrieval systems that can cross language boundaries. Kannada is used across educational resources, public information, digital documents and regional knowledge collections. However, a user who formulates a query in English may have difficulty finding relevant Kannada content when the retrieval process is based mainly on exact lexical matching. Such matching is particularly restrictive when the query and document use different languages or when expressions with similar meanings use different words.

Cross-Language Information Retrieval enables a query written in one language to locate information stored in another. In CrossLinguaRAG, multilingual Sentence-BERT [1], [2] maps English and Kannada text into a shared representation space. The retrieval engine does not depend on a single similarity measure; instead, it combines semantic similarity with lexical and keyword evidence. The retrieved passages are then provided to a Retrieval-Augmented Generation (RAG) component [3] so that the generated English answer is based on identifiable Kannada evidence.

A. Objectives

- Develop a cross-language retrieval system for Kannada document collections.
- Support English and Kannada queries over Kannada PDF, DOCX and TXT files.
- Represent document passages and queries using multilingual Sentence-BERT embeddings.
- Combine semantic, TF-IDF lexical and keyword-based retrieval signals.
- Return transparent Kannada evidence with English translation.
- Generate evidence-grounded English answers through RAG.

II. METHODOLOGY

The system is implemented in two main stages. The first stage builds the document knowledge base, and the second stage handles the user query and generates an evidence-based answer. During knowledge-base construction, Kannada documents are extracted, segmented, encoded and stored. During a search operation, an English or Kannada query is normalized, embedded and compared with indexed passages. The candidate passages are ranked using a hybrid score, translated when necessary, and supplied to the generation component (Fig. 1).

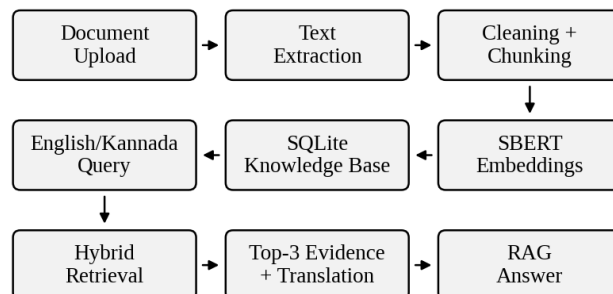


Fig. 1. End-to-end processing pipeline of CrossLinguaRAG.

A. Input Collection and Document Processing

The application accepts Kannada PDF, DOCX and UTF-8 TXT documents. The backend extracts textual content and performs language-content validation before cleaning the text.

The cleaned material is divided into manageable overlapping chunks. Chunking prevents the retrieval stage from treating an entire document as one large unit and instead provides focused passages that can be ranked independently.

B. Embedding and Retrieval

Every document chunk is transformed into a dense multilingual representation with the pretrained paraphrase-multilingual-MiniLM-L12-v2 Sentence-BERT model [1], [2], [7]. The user's query is encoded with the same model so that English and Kannada representations can be compared in a common semantic space. Cosine similarity provides the semantic component of retrieval. TF-IDF [4], [5] contributes a lexical signal, while keyword matching identifies explicit query-term overlap. These signals are combined to form a hybrid relevance score, after which the highest-ranked passages are selected.

C. RAG Answer Generation

The selected evidence is passed to the Gemini generation component [6] together with the original query. The generation instruction directs the component to base its response on the retrieved material and return the answer in English. The original Kannada passages and their English translations remain available to the user, allowing the generated response to be checked against the retrieved evidence.

III. MODELING AND ANALYSIS

The main functional modules of the implemented system are summarized in Table I.

TABLE I
FUNCTIONAL MODULES OF THE SYSTEM

No.	Component	Functional Description
1	Document Input	Accepts Kannada PDF, DOCX and TXT files.
2	Preprocessing	Extracts, cleans and chunks document text.
3	Embedding	Creates multilingual Sentence-BERT representations.
4	Semantic Retrieval	Uses cosine similarity between query and passage embeddings.
5	Lexical Retrieval	Uses TF-IDF to capture word-level overlap.
6	Keyword Matching	Adds explicit matching of important query terms.
7	Hybrid Ranking	Combines semantic, lexical and keyword signals.
8	Translation	Provides English translations of Kannada evidence.
9	RAG	Generates an evidence-grounded English response.
10	Database	Stores indexed chunks and associated metadata.

IV. PROBLEM STATEMENT

Regional-language document collections contain valuable information, yet conventional retrieval systems frequently assume that the query and document share a language. For Kannada collections, an English-speaking user may therefore be unable to formulate an effective Kannada keyword query. Exact matching also misses semantically related expressions when different words are used. The required system must identify the semantic relationship between English and Kannada text, rank useful evidence and present the retrieved information in an accessible form.

V. PROPOSED SYSTEM

CrossLinguaRAG is an end-to-end web-based CLIR and RAG system. Its architecture keeps document indexing, retrieval, translation and generation as separate stages, as shown in Fig. 2. This separation makes the evidence produced by retrieval visible and ensures that the generation stage operates after relevant passages have been identified.

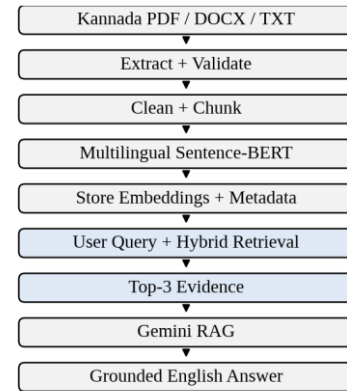


Fig. 2. Proposed CrossLinguaRAG processing architecture.

A. Query Processing

The query is normalized and encoded with the same multilingual model used for document chunks. For an English query, this shared representation permits direct comparison with Kannada passage embeddings. The retrieval engine calculates semantic, lexical and keyword signals and combines them into a relevance score. The system returns the top three evidence passages together with confidence-related retrieval information.

B. Evidence Translation and Grounded Generation

Retrieved Kannada passages are translated into English for accessibility. The RAG component receives the retrieved evidence and the original query and produces an English response. Because the response is based on retrieved evidence, the source passages remain available for inspection.

VI. EXISTING SYSTEM

Existing approaches show the following limitations:

- Keyword-based retrieval depends largely on lexical overlap between the query and document.
- Conventional same-language retrieval does not directly solve cross-language search.
- Pure semantic retrieval may be less discriminative for very short or ambiguous queries.
- Generation systems without visible evidence may make source verification difficult.

CrossLinguaRAG addresses these limitations through multilingual embeddings, hybrid ranking, visible evidence and grounded RAG.

VII. RESULTS AND DISCUSSION

To check the retrieval performance, we used a manually annotated set of 20 English and Kannada queries. For each query, the top three retrieved passages were inspected using relevance grades of 0 (not relevant), 1 (partially relevant) and 2 (highly relevant). Retrieval performance was summarized with Recall@1, Recall@3, MRR and nDCG@3 [4], as reported in

Table II. Since the test set contains only 20 queries, these results describe the current prototype and should not be treated as a complete evaluation of Kannada information needs.

TABLE II
 RETRIEVAL PERFORMANCE ON THE 20-QUERY TEST SET

Metric	Overall	English	Kannada
Recall@1	80.0%	60.0%	100.0%
Recall@3	90.0%	80.0%	100.0%
MRR	0.8333	0.6667	1.0000
nDCG@3*	0.8695	0.7470	0.9920

*nDCG@3 uses graded relevance; the query with no relevant result is treated as zero in the overall average.

A. Interpretation

In this test collection, Kannada queries achieved higher first-rank retrieval performance, whereas English queries showed greater variation. This difference is consistent with the current retrieval configuration and the limited size of the evaluation collection. Short cross-language queries can be less discriminative in embedding space. The hybrid method supplements semantic similarity with lexical and keyword evidence, although the combination does not remove every ambiguity.

B. Qualitative Retrieval Examples

TABLE III
 SAMPLE QUERIES AND RETRIEVAL OUTCOMES

Query	Observation	Outcome
Districts famous for coffee?	Kannada evidence identifies Kodagu and Chikkamagaluru.	Relevant
Festival famous in Karnataka?	Mysuru Dasara appears among retrieved evidence.	Grounded RAG
ಬೆಂಗಳೂರು ಸರ್ಕಾರದ ರಾಜಧಾನಿಯೇ?	The retrieved passage states that Bengaluru is the state capital.	Direct

C. RAG Output

For a query concerning Karnataka's coffee-growing districts, the system retrieved a Kannada passage identifying Kodagu and Chikkamagaluru. The RAG stage produced the English response that these two districts are famous for coffee cultivation in Karnataka. The example demonstrates the intended separation between evidence retrieval and response generation.

D. Limitations and Discussion

- The evaluation dataset contains only 20 queries and does not represent the full range of Kannada information needs.
- Short or broad queries can produce passages that are semantically related but insufficiently specific.
- Evidence quality depends on document extraction, chunking and the indexed collection.
- Translation quality may vary with the source text and translation service.
- If the correct evidence is not retrieved, the generation module cannot reliably recover the missing information.

VIII. CONCLUSION

CrossLinguaRAG presents a cross-language information retrieval architecture for accessing Kannada document content through English or Kannada queries. The system combines

document processing, multilingual Sentence-BERT embeddings, semantic similarity, TF-IDF, keyword matching, translation and Retrieval-Augmented Generation in a single web-based workflow. Unlike a generation-only question-answering interface, the architecture exposes the retrieved Kannada evidence so that users can inspect the material supporting the final English response.

The results show that the prototype can retrieve useful Kannada passages for both English and Kannada queries and use those passages to produce English answers grounded in the retrieved content. As a next step, the project can be extended by a larger manually annotated Kannada CLIR benchmark, stronger query-passage datasets, systematic comparison with additional retrieval baselines, improved translation quality and more extensive evaluation of grounded answer generation.

IX. IMPLEMENTATION SUMMARY

The implemented workflow is illustrated in Fig. 3, and the technology stack is listed in Table IV.

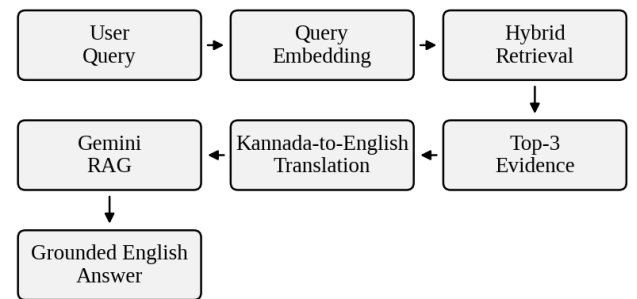


Fig. 3. Integrated CLIR and grounded-answer workflow.

TABLE IV
 IMPLEMENTATION STACK

Layer	Implementation
Backend	Python and Flask [8], [9]
NLP / Retrieval	Multilingual Sentence-BERT, cosine similarity, TF-IDF and keyword matching
Supported documents	PDF, DOCX and TXT
Database	SQLite [10]
Generation	Gemini API [6]
Frontend	HTML, CSS and JavaScript
Embedding model	paraphrase-multilingual-MiniLM-L12-v2

A. Knowledge-Base and Search Operation

During indexing, uploaded Kannada documents are processed and divided into searchable passages. The embedding model converts each passage into a multilingual representation, while document metadata is retained in the SQLite knowledge base. At search time, the query is encoded with the same pretrained model. The retrieval engine compares the query with stored passages and combines semantic similarity, TF-IDF and keyword evidence before returning the highest-ranked results.

B. Evidence-Grounded Generation

The RAG stage is positioned after retrieval rather than being used as the primary search mechanism. Retrieved passages are supplied to the generation component together with the user's query. The intended response is therefore based on the selected

evidence, while the Kannada source material and translation remain visible for verification. This design separates the responsibilities of retrieval and generation.

C. Model and Training Note

The Sentence-BERT model used in this project is a pretrained model; the paper does not claim that it was fine-tuned for the project. Uploaded documents are indexed into the application's knowledge base and do not retrain the embedding model. The system therefore uses the pretrained multilingual representation as the semantic foundation for cross-language comparison.

D. Future Development

- Build a larger manually annotated Kannada CLIR benchmark.
- Expand the variety and difficulty of English-to-Kannada and Kannada-to-Kannada queries.
- Compare the hybrid retriever with additional retrieval baselines.
- Improve translation quality and evaluate it systematically.
- Conduct broader evaluation of evidence-grounded answer generation.

ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering, MRIT, and Prof. Kiran B for guidance and support throughout this project.

REFERENCES

- [1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019.
- [2] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in Proc. EMNLP, 2020.
- [3] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020.
- [4] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [5] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," J. Documentation, vol. 28, no. 1, pp. 11–21, 1972.
- [6] Google, "Gemini API documentation," Google AI, 2024.
- [7] "Sentence Transformers documentation," SBERT.net, 2024.
- [8] "Flask documentation," Pallets Projects, 2024.
- [9] Python Software Foundation, "Python documentation," 2024.
- [10] "SQLite documentation," SQLite Consortium, 2024.