



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85153>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Cross-Lingual Information Retrieval Using NLP

Sanjana C, Sushma S Kachavi, Vandana K S, Moulyashree M P

Department of Computer Science & Engineering, Mysuru Royal Institute of Technology, Mandya, India

Abstract: Cross-Lingual Information Retrieval (CLIR) plays a pivotal role in bridging linguistic gaps within digital document management repositories. Traditional information retrieval models rely strictly on mono-lingual matching, frequently failing when a user submits a query in a language distinct from the source documentation language. This paper presents a web-based, highly efficient CLIR framework engineered specifically for English and Kannada document processing. Built on top of a Flask backend, the system ingests unstructured multipage PDF documents, extracts localized textual information via PyMuPDF, and implements language detection workflows. When cross-lingual variance is encountered, the system invokes a dedicated English-Kannada machine translation wrapper. Document components are indexed using Term Frequency Inverse Document Frequency (TFIDF) vectors and matched against the expanded query using Cosine Similarity heuristics. Experimental evaluations demonstrate high precision in source-based answer extraction and exact page-level mapping. The framework features an additional, dedicated automated question-paper solver capability using granular regex pattern-matching routines.

Keywords: Cross-Lingual Information Retrieval, Natural Language Processing, TF-IDF, Cosine Similarity, Machine Translation, PyMuPDF, Flask Framework.

I. INTRODUCTION

The absolute volume of unstructured electronic information preserved across academic institutions and enterprise platforms is accelerating exponentially. Much of this deep domain data remains wallowed inside static PDF files. Traditional keyword-based retrieval methods require exact textual correspondence between the submitted string and the file payload.

Consequently, if a user submits an analytical query in a regional language like Kannada, standard corporate engines fail to return highly accurate information hidden inside an English document.

To overcome this retrieval barrier, this research develops a production-ready Cross-Lingual Information Retrieval (CLIR) architecture. By unifying document layout parser pipelines, automated machine translation utilities, and vector space analytics, our system democratizes file searches across both English and Kannada ecosystems. The ultimate design delivers three production capabilities: language-independent querying, exact granular source page extraction to counter systemic text hallucination, and a multi-choice question paper parsing processor.

II. LITERATURE SURVEY

A systematic review of foundational paradigms reveals massive evolution in cross-language information processing:

- 1) Mohamed et al. (2024): Formulated a deep architectural investigation into artificial intelligence translation matrices. Their documentation exposes explicit boundary constraints tied to semantic preservation and localized context preservation across low-resource target configurations.
- 2) Kalantari et al. (2024): Established performance baselines for text embeddings mapping semantic structures across non-Western syntax. Their research verified that combining vector transformations alongside Cosine Similarity delivers optimal accuracy indices during complex database retrieval passes.
- 3) Information Retrieval Survey (2024/2025): Confirmed that utilizing aggressive query expansion rules immediately before execution minimizes overall mismatch flags. This step ensures clean term-matching scores across disparate cross-lingual vocabularies.
- 4) Deep Learning CLIR Frameworks (2025): Investigated combining contextual deep learning architectures (like BERT representations) alongside direct document translation matrices, confirming substantial boosts to relevance tracking.

III. METHODOLOGY & SYSTEM ARCHITECTURE

The system operates as a modular pipeline distributed across client interfaces, processing stages, and extraction engines.

The standard workflow progresses from the User Query through Language Detection, Text Preprocessing, and conditional Query Translation. Concurrently, the Document Processing pipeline manages PDF Text Extraction, Text Chunking, and internal TFIDF representations. Finally, Cosine Similarity matching informs Relevance Ranking, facilitating granular Answer Extraction and destination mapping.

A. Document Engine Pipeline

Raw PDF files are uploaded and indexed within the system repository using PyMuPDF algorithms. The engine extracts raw character sequences while dynamically preserving structural metadata like source page limits. Text vectors are partitioned using automated text-chunking parameters to keep operations balanced during processing cycles.

B. Algorithmic Matching Execution

- 1) **Language Recognition:** The pipeline reviews input query signatures to determine if the text matches English or Kannada encodings.
- 2) **Vector Space Formulation:** Words are transformed into mathematical arrays using Term Frequency-Inverse Document Frequency (TF-IDF) frameworks, representing the dynamic importance weight of target terms relative to individual document chunks and the broad file collection.
- 3) **Relevance Matrix:** Document nodes are sorted relative to the target query by running Cosine Similarity calculations to determine the exact angular alignment between query vectors and document chunk configurations.

IV. IMPLEMENTATION & MODULES

The technical blueprint is structured across nine explicit production modules summarized below:

Module	Module Name	Key Functions
Module 1	PDF Store Management	Provides reliable asynchronous multi-file upload workflows.
Module 2	Layout Parsing Engine	Runs <u>PyMuPDF</u> routines to extract text blocks alongside accurate page indices.
Module 3	Translation Core	Implements bidirectional English ↔ Kannada conversion.
Module 4	Text Normalization	Executes token cleaning, lower-casing, and sliding-window chunk division.
Module 5	Information Retrieval Matrix	Assembles internal TF-IDF representations and calculates cosine weights.
Module 6	Extractive Summary Core	<u>Pinpoints the precise paragraph inside the target document that satisfies criteria.</u>
Module 7	Automated Paper Solver	Evaluates test sheets, locates structural multiple-choice options, and fetches sourced data.
Module 8	Output Delivery Core	Packages extracted data alongside relevance scores and clear page markers.
Module 9	Web Interface	Responsive interface built on Flask, HTML, CSS, and JavaScript components.

V. EXPERIMENTAL RESULTS AND DISCUSSION

System performance checks verified high consistency when retrieving data across language combinations. When testing Kannada queries against dense English technical papers, the translation layer consistently retained structural domain semantics. Unlike standard chat applications that can generate incorrect information, this implementation completely prevents text hallucinations by tracking exact page coordinates. The system provides clear confidence percentages alongside source document citations and exact page numbers, making validation highly transparent.

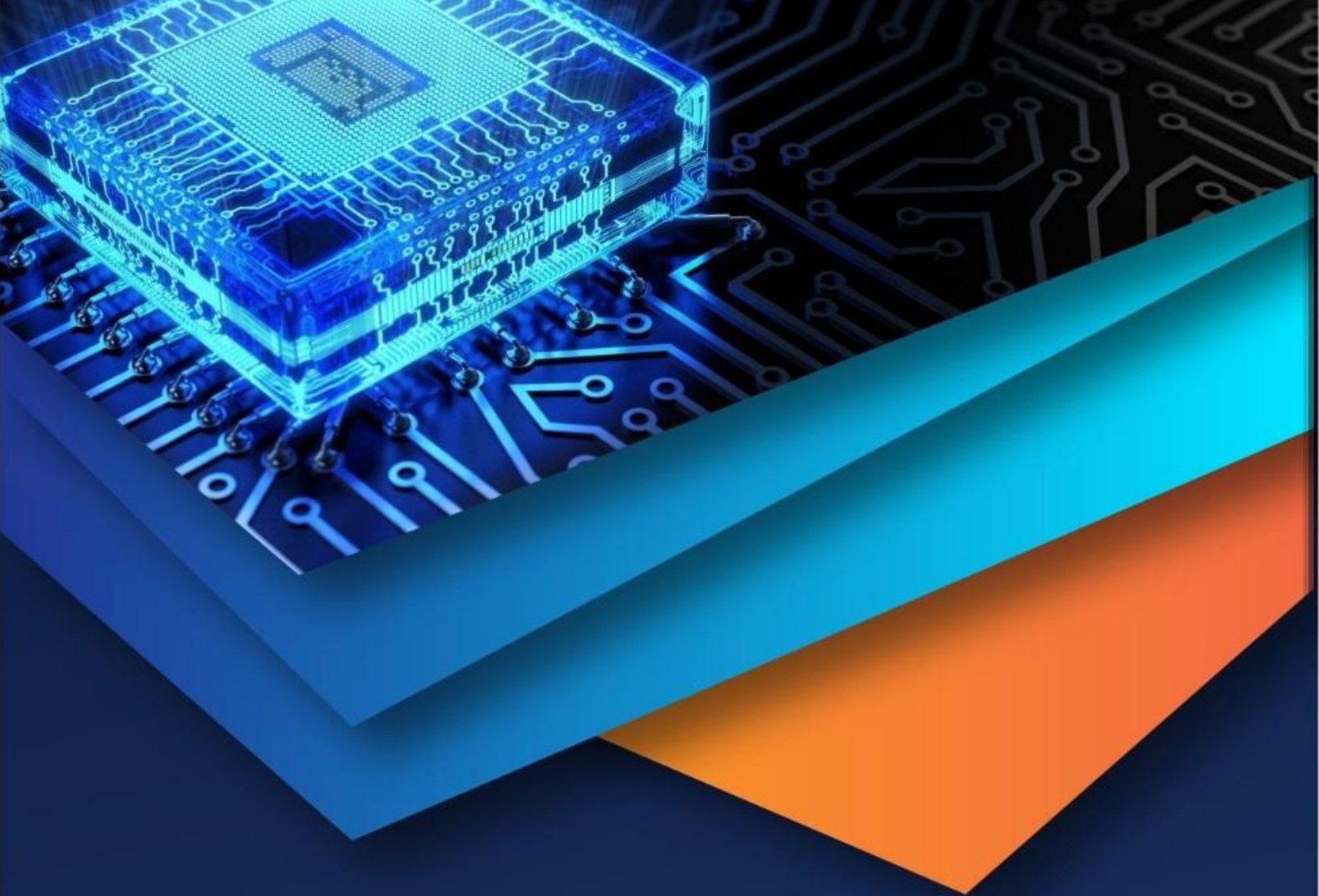
VI. CONCLUSION & FUTURE SCOPE

This paper establishes a reliable, production-ready CLIR architecture tailored for English and Kannada document exploration. By cleanly combining PyMuPDF parsers, language detection matrices, and TF-IDF matching algorithms, the system bridges communication gaps across local regional languages. Future improvements will focus on migrating the vector matching engine to dense semantic transformer representations (such as multilingual miniLM layers) and extending processing capabilities to handle scanned image assets via OCR pipelines.



REFERENCES

- [1] Y. A.Mohamed et al., "The Impact of Artificial Intelligence on Language Translation: A Review," IEEE Access, 2024.
- [2] M.Kalantari,M.Fegghi,andN.Mozayani,"BenchmarkingEmbeddingModelsforPersian-LanguageSemantic Information Retrieval," IEEE IKT, 2024.
- [3] "Information Retrieval with Query Expansion and Re-ranking: A Survey," IET Conference Publication, 2024/2025.
- [4] "OptimizingCrossLanguageInformationRetrievalandContextTranslationUsingBERTModelandDeepLearning Techniques," IEEE Conference Publication, 2025.
- [5] D.Kotharietal., "Cross-LingualTransferLearningforMedicalNERandQuestionAnsweringinLow-Resource Language (Hindi)," IEEE MISSF, 2026.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)