ISSN No:-2456-2165

Bridging Language Gaps in Health Information Access: Konkani-English CLIR System for Medical Knowledge

Gaurav Mahajan
Dept. of Computer Science
PES University
Bengaluru, India

Abstract:- This paper addresses the challenges posed by linguistic diversity in terms of medical information by introducing a Cross-Language Information Retrieval System attuned to the needs of Konkani language information seekers. The proposed system leverages Konkani queries entered by the user, translates them to English, and retrieves the documents using a thesaurus-based approach. Various strategies also have been considered to address the challenges posed by the source language – Konkani which is a minority language spoken in the Indian subcontinent. The proposed approach showcases the potential of combining language technology, information retrieval, and medical domain expertise to bridge linguistic barriers. As healthcare information remains a critical societal need, this work holds promise in facilitating equitable access to medical knowledge.

Keywords:- *Information Retrieval, Medical Knowledge, Konkani, Minority Language, Health, Thesaurus-based Approach.*

I. INTRODUCTION

In today's digital age, the rapid growth of data and information has completely revolutionized how internet users access and retrieve information across a plethora of domains. In the healthcare sector, Information Retrieval (IR) plays a pivotal role in providing accurate and timely information – crucial for both health professionals and the general public. The field of Information Retrieval plays an important role within the field of computer science ensuring that relevant and reliable information is provided to those who seek it.

Also known as Search Technology, IR encompasses a diverse range of techniques, algorithms, and methodologies that aim at efficiently locating and delivering relevant information from vast repositories, corpora, and datasets. While traditional IR systems are aimed at providing information in a single language, the challenges posed by multilingualism are undeniable. The World Wide Web has facilitated the dissemination and propagation of knowledge and information across borders, languages, and cultures, and thus highlighting the necessity of overcoming language barriers for effective knowledge access.

The diversity of languages spoken across the globe has spurred the development of multilingual IR systems, collectively referred to as Cross-Language Information Retrieval (CLIR). These systems enable any user to submit queries in one language and retrieve relevant documents in another. One of the most compelling aspects of CLIR is its possibilities to empower underrepresented content. Many languages have limited digital content available in them, thus making them susceptible to being marginalized in the digital age. CLIR has immense potential to democratize information access, especially for underrepresented languages and minority communities.

One such language that deserves attention is Konkani. Konkani is one of the 22 scheduled languages of India. Spoken mainly across the Western Coastal region which spans the states of Maharashtra, Goa, Karnataka, and Kerala, and Union Territories like Daman and Diu, and Dadra and Nagar Haveli. However, it is the official language of Goa and is spoken by around 2.5 million people. There is no specific or exclusive script for the language but written in Devanagari, Roman, and Kannada scripts. When a language spans across multiple regions, there is a strong presence of dialects. Some of the major dialects are Goan Konkani, Canara Konkani, GSB Konkani, Maharashtrian Konkani, and Mangalorean Konkani.

Konkani, a language of rich cultural heritage, is a language that requires special attention in terms of research. It serves as a linguistic bridge, reflecting a linguistic amalgamation of elements from Sanskrit, Marathi, Portuguese, and Kannada languages. The characteristics of this language, including its script and vocabulary variations, pose unique challenges in designing effective IR systems. In this paper, the focus is on Goan Konkani in Devanagari script, which is considered a standard representation of the Konkani language.

In this paper, I contribute to the advancement of CLIR by introducing a specialized system tailored for health information retrieval in the Konkani language. Leveraging cutting–edge translation techniques and novel retrieval strategies, my approach seeks to streamline access to health-related information for Konkani speakers which also justifies that a majority population is rural with no immediate access to health services.

This paper provides the design and implementation of a novel CLIR which includes query pre-processing to enhance the retrieval of documents, a translation module – that converts Konkani queries to English, construction of an inverted index – to ensure swift access retrieval and evaluation in the form of relevance feedback. By concentrating on the unique linguistic landscape of Konkani, I extend the ambit of CLIR to serve a previously underserved linguistic community.

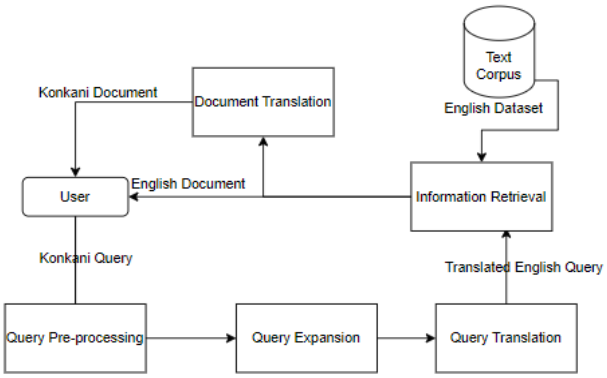

Fig. 1. The proposed Konkani Health CLIR system

II. RELATED WORK

The work of Mallamma V. Reddy and M. Hanumanthappa [1] makes significant contributions to the field of CLIR. Their research focuses on the issue of translating native language queries – especially specified in Kannada and Telugu into English while still preserving the relevance of retrieved details. With the use of bilingual dictionaries and transliteration for words that are absent in the dictionary, they use a query translation approach. A translation disambiguation module based on co-occurrence statistics is also included in their system. They achieved a commendable gain in retrieval performance for Kannada and Telugu documents compared to traditional monolingual IR systems.

The paper authored by P. Sujatha, P. Dhavachelvan, and V. Narasimhulu, [2] presents a CLIR system that translates English queries into Telugu and Tamil to retrieve documents from the computer science domain. Their system utilizes bilingual dictionaries and transliteration techniques. The paper also provides insights into research on CLIR systems for Indian languages.

In their work, K. Lokhande and D. Tayade [3] address the challenge of retrieving documents in Marathi when users input queries in English. The authors suggest an English-Marathi CLIR system that makes use of WordNet for query expansion, pre-processing, and translation. Tokenizing English queries, expanding them with WordNet to produce related queries, and then translating the improved English query into Marathi using Google Translator are the steps in the procedure. Following that, Marathi documents based on term matching are retrieved using the translated Marathi query. The FIRE 2010 dataset for Marathi was used by the authors to evaluate the system's performance using precision and recall as evaluation metrics. The findings showed that the system had good precision and recall values, which made it a promising CLIR solution. The paper concludes by emphasizing the value of CLIR in removing language barriers on the internet and offering suggestions for future research areas, like creating linguistic resources.

The paper by Sakthi Vel S. and Dr.Priyar [4] discusses various text pre-processing techniques such as dataset creation, tokenization, noise removal, stop word removal, stemming, lemmatization, and term weighting, specifically for Tamil and Malayalam languages. Their findings emphasize the importance of text pre-processing in CLIR systems. They study in detail the impact of different stemming techniques and compare different text mining and pre–processing techniques. They provide an algorithm outlining the text pre-processing steps for the collected documents, including tokenization, URL removal, symbol and punctuation removal, stop word removal, stemming, lemmatization, and term weighting using TF-IDF.

In their paper, Paheli Bhattacharya, Pawan Goyal, and Sudeshna Sarkar [5] suggest a novel approach to word embeddings for Cross-Language Information Retrieval (CLIR). The performance of CLIR is significantly enhanced by this method, especially for Hindi to English translations. They present a method that makes use of word embeddings to extract contextual information from source-language words and locate translations in the target language that have similar contexts. This technique relies solely on a dictionary with word translation pairs for training the word vectors, eliminating the need for aligned corpora, a common problem in CLIR. The FIRE 2008 and 2012 datasets for Hindi to English CLIR are used by the authors in their experiments. Their word embedding-based method outperforms the fundamental dictionary-based approach by a sizeable margin of 70%. The hybrid approach outperforms the standard dictionary-based method by 77% when word embeddings and the dictionary are combined. Moreover, it shows a 15% improvement over the English-only baseline when using dictionary and Google Translate translations.

Jacques Savoy, Ljiljana Dolamic, and Mitra Akasereh's paper [6] focuses on information retrieval strategies for Hindi, Bengali, and Marathi. They investigate indexing and search methods such as aggressive stemming, n-gram language-independent approaches, and trunc-n indexing. The study employs various information retrieval models and discovers that those based on the Divergence from Randomness (DFR) paradigm consistently outperform the others. Furthermore, aggressive stemming and trunc-n indexing boost retrieval efficiency. After blind-query expansion, using the Z-score as a data fusion operator improves the Mean Average Precision (MAP) in merged runs over individual systems.

The purpose of K.M. Chaman Kumar *et al's* [7] paper is to investigate machine translation methods for translating Konkani to English. It covers both traditional approaches, such as rule-based and statistical machine translation (SMT), and modern techniques, such as neural machine translation

(NMT) and hybrid-based machine translation (HMT). For Konkani translation, the authors propose using Statistical Machine Translation (SMT) with tools such as MOSES and GIZA++. They go over different stages of the translation process, such as tokenization and language modeling. They also discuss the use of Neural Machine Translation (NMT) and its benefits and challenges, as well as its architecture and tools such as OpenNMT. The paper discusses the relevance and accuracy of Hybrid Machine Translation (HMT), which combines various techniques.

Savita Mayanale and S. S. Pawar [8] discuss the development of a Marathi–English CLIR system considering challenges such as lexical ambiguity and query term mismatch. A query translation approach is used in the proposed system, in which Marathi queries are translated into English using a bilingual dictionary. A transliteration method is used if a translation is not available. A Word Sense Disambiguation (WSD) module is also included to improve retrieval accuracy. To find synonyms for correctly sensed words, this module employs an unsupervised corpus-based method and an English dictionary. The experimental results show that the proposed method improves the performance of Marathi-English CLIR, particularly when detailed queries are used. On the FIRE 2011 dataset, the system demonstrated high precision levels.

The paper by Parin Chheda, Manaal Faruqui, and Pabitra Mitra [9] addresses the issue of handling out-of-vocabulary (OOV) words in Indian-language to English Cross-Lingual Information Retrieval (CLIR). OOV words are those that are not present in translation resources, causing retrieval issues. The paper proposes a solution that includes word segmentation and approximate string matching with Soundex and Levenshtein distance. In comparison to a baseline CLIR system, evaluations on three Indian languages show an average 5.8% improvement in retrieval accuracy. Monolingual, approximate string matching, word segmentation, and maximal-prefix word segmentation are among the models proposed. This study demonstrates a promising method for improving CLIR in Indian languages plagued by OOV words.

The work by Teklay Birhane and Birhanu Hailu [10] describes the creation of an Information Retrieval (IR) system for Tigrigna textual documents to meet the needs of Tigrigna language users looking for relevant information online. The system is divided into two parts: indexing, which uses keyword extraction to organize Tigrigna documents, and searching, which uses the Vector Space Model (VSM) for similarity calculations and ranking. For Tigrigna documents, the study achieves 70% precision and 84% recall. Their work inspired me to take up Konkani as a source language.

III. DATASET

The dataset for the system is constructed using health-related articles from Wikipedia. A corpus is created by web crawling web pages and stored in a text file format. It is also necessary to handle the HTTP errors as well which are taken care of by adding conditional checking. The corpus consists of Wikipedia articles related to health in general, diseases, and other miscellaneous medical domain-related documents. There are a total of 2287 documents.

IV. IMPLEMENTATION

This section specifies the implementation details of developing a CLIR system attuned to handle health-related queries in Konkani.

A. Text Preprocessing for Documents

One of the primordial steps is to preprocess text present in the documents to ensure consistency in the data and relevance during retrieval. The steps are as follows:

- *Case-folding:* All the contents are converted into lowercase. Case folding is one of the most basic and effective pre-processing techniques and helps overcome irregularities and discrepancies in the text.
- *Tokenization:* The text is broken down into simpler words or terms. It helps create a better-inverted index.
- *Stemming:* Since there are a lot of words that are inflections of the common word stem, stemming is used to represent such words.

B. Inverted Index

The efficacy of an IR system lies in the construction of the inverted index. An inverted index is a data structure that maps a unique word to a list of documents that it is present in. Since the word-document mappings are stored in a sorted manner, it reduces the time taken to retrieve the documents as there is no scanning of the entire collection. Post-pre-processing, the implementation involves the construction of an inverted index.

C. Query Translation

The implementation of the system proceeds with incorporating a translation module that will convert the queries submitted by the user in Konkani to English. This step is necessary as the results of the translation determine the success of the system. The translation module is designed using the IndicTrans model, an open-source translation model developed using a multilingual NMT model. It is one of the few models that gives accurate translation and transliteration of Konkani text in Devanagari script, alongside Google Translate.

D. Query Processing and Expansion

After translating the queries, processing of queries is done to identify relevant keywords. Queries are processed the same way as text in documents. Then, the inverted index performs a lookup for each term in the query and consolidates documents containing those terms.

Since the query is processed, the original scope or intent should not be lost. Hence, the process of query expansion plays a role in widening the scope of user queries. This also improves search performance. The procedure entails adding more related words, synonyms, or conceptually comparable terms to the initial query using lexical resources like thesauri. The system aims to return a more thorough and relevant

search by including these expanded terms in the query. When dealing with problems like polysemy and synonymy, where a single word can have several meanings or different words can represent the same concept, query expansion is especially helpful. Information retrieval systems' search accuracy and user satisfaction are improved through the inclusion of more terms, making the query more robust and contextually rich.

E. Ranking

After retrieval of relevant documents, the retrieved documents are ranked based on their relevance with each query term. Ranking is done using a popular algorithm which is TF-IDF. This handy algorithm uses the frequency of words to measure the weightage or relevance of words to a document. The ranking module follows a Top-K approach where K=10, in other words, the top 10 ranking documents are shown to the user.

V. CHALLENGES

The design and implementation of a Cross-Language Information Retrieval system come with a lot of challenges. When developing such a system using a minor language like Konkani, the challenges are as follows:

(a) Lack of Resources: One of the primary challenges faced is the scarcity of digital content, translation models, and dictionaries. The creation of accurate models is still a challenge.

(b) Linguistic Variation and Lack of Standardization: As discussed earlier, the existence of various dialects and the lack of standard reference resources, it difficult to accommodate user requests.

(c) Translation Quality: CLIR systems rely on translation models to bridge the gap between languages. For Konkani, which does not have well-established machine translation models, achieving accurate translations can be difficult.

(d) Cross-Language Relevance: Ensuring that the retrieved documents are relevant to the user's query and not just translations of keywords is a significant challenge. CLIR systems should consider semantic equivalence and context.

VI. RESULTS

Since there are no established evaluation metrics that correlate Konkani and English document retrieval, the system uses relevant feedback from the user to improve the performance. The results could be better represented graphically.

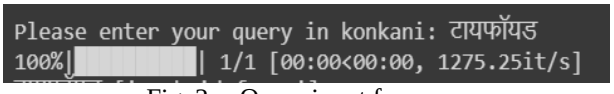

Fig. 2 a. Query input from use

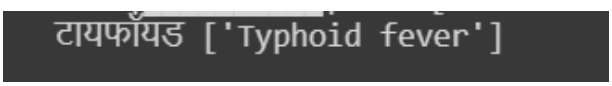

Fig. 2 b. Query translated to English

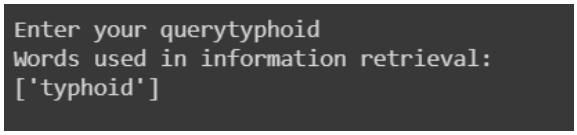

Fig. 2 c. Query tokenized

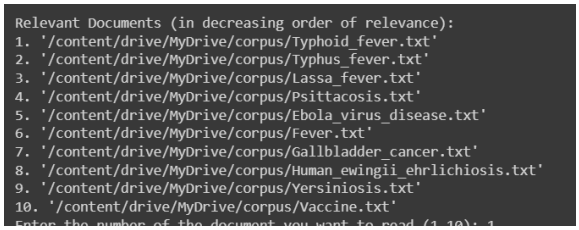

Fig. 2 d. Top 10 documents retrieved

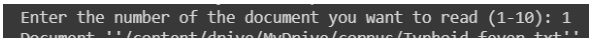

Fig. 2 e. The user selects the document retrieved

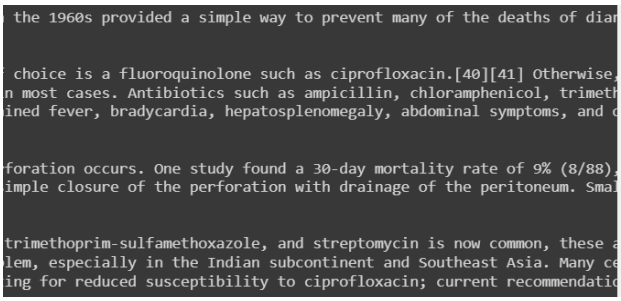

Fig. 2 f. Document displayed

VII. CONCLUSION

To conclude, this paper brings in a dedicated Cross-Language Information Retrieval system targeted towards the health domain. The system can efficiently accept queries in Konkani from the end-user and translate them into English. The system also employs techniques that act as a base for any IR system such as the construction of an inverted index, and thesaurus-based information retrieval technique. The system fulfills its target to initiate ground-breaking research work on one of the underserved Indian languages like Konkani, thus breaking linguistic barriers and helping to bridge the gaps between health professionals and native speakers.

During the development, the challenges have inspired me to take up opportunities to enable effective IR in a language as diverse and rich as Konkani. This paper sheds light on the complexities of a CLIR in a language with limited digital resources, different scripts, and varying dialects. Since this system has made strides in developing a foundation for Konkani CLIR, there are still a lot of hurdles yet to overcome.

VIII. FUTURE WORK

Looking ahead, the absence of a standardized Konkani digital thesaurus and dictionary should be considered to develop better translation models. Having constructed a rudimentary CLIR system, this paper paves the way for the creation of larger, more comprehensive datasets. I also identified the need for a Konkani–specific stemmer and stop-

word list, essential components for improving the precision of the system.

It is also evident from the literature survey that realizing the full potential of a Konkani CLIR requires a collective effort. Collaboration between linguists, technologists, and researchers will be pivotal in addressing the complexities of the Konkani script and dialects. Moreover, the development of more advanced language models, trained specifically for Konkani, holds great promise.

Considering the health CLIR for Konkani, the future lies in developing advanced features like:

- *Semantic Matching:* This NLP technique could be applied to measure the degree of similarity between words in terms of their semantics or meaning. This would enable a context–aware processing.

- *Real-time Data:* Involve and find out various APIs and data from health institutions that could be leveraged for extended use cases like predictions and diagnosis.

- *Interface for User Interaction:* A user-friendly interface can be integrated to facilitate hassle-free access, and feedback and ensure relevant information.

- *Machine Learning and Recommendation Algorithms:* To make the systems smarter more powerful and more user-centric and personalized, machine learning algorithms for translation and recommendation algorithms for user experience can be leveraged.

- *Handling the Roman and Kannada scripts:* The present system deals with Konkani written in Devanagari script. Various transliteration mechanisms can be developed and used to cater to a larger diaspora of Konkani speakers.

In the spirit of continued exploration and innovation, I encourage future researchers to build upon my work. Together, we can forge a system where not only Konkani speakers but those of all the languages in the world will be able to preserve and celebrate the linguistic tapestry of our world.

REFERENCES

[1]. M. V. Reddy et al., "Kannada and Telugu Native Languages to English Cross Language Information Retrieval," in International Journal of Computer Science and Information Technologies, vol. 2, no. 5, pp. 1876-1880, 2011.

[2]. P. Sujatha, P. Dhavachelvan, and V. Narasimhulu, "Evaluation of English-Telugu and English-Tamil Cross-Language Information Retrieval System using Dictionary Based Query Translation Method," International Journal of Computer Science and Information Security, vol. 8, no. 2, pp. 314-319, 2010.

[3]. K. Lokhande and D. Tayade, "English-Marathi Cross-Language Information Retrieval System," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 8, pp. 112-116, August 2017. [Online]. Available: ISSN: 2277-128X, DOI: 10.23956/ijarcsse/V7I8/0127.

[4]. S. V. S and P. R, "Text Pre-Processing Methods on Cross-Language Information Retrieval," 2022 International Conference on Connected Systems & Intelligence (CSI), Trivandrum, India, pp. 1-5, 2022. DOI: 10.1109/CSI54720.2022.9923952.

[5]. P. Bhattacharya, P. Goyal, and S. Sarkar, "Using word embeddings for query translation for Hindi to English cross-language information retrieval," Computacion y Sistemas, vol. 20, no. 3, pp. 435–447, 2016. DOI: 10.48550/arXiv.1608.01561

[6]. J. Savoy, L. Dolamic, and M. Akasereh, "Information Retrieval with Hindi, Bengali, and Marathi Languages: Evaluation and Analysis," in Multilingual Information Access in South Asian Languages, P. Majumder, M. Mitra, P. Bhattacharyya, L. V. Subramaniam, D. Contractor, and P. Rosso, Eds. Springer, Berlin, Heidelberg, 2013, vol. 7536, DOI: 10.1007/978-3-642-40087-2_30.

[7]. K. M. Chaman Kumar, S. Aswale, P. Shetgaonkar, V. Pawar, D. Kale, and S. Kamat, "A Survey of Machine Translation Approaches for Konkani to English," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, pp. 1-6, 2020. DOI: 10.1109/ic-ETITE47903.2020.110.

[8]. Savita C. Mayanale and Ms. S. S. Pawar, "Marathi-English CLIR using detailed user query and unsupervised corpus-based WSD," *Int. Journal of Engineering Research and Applications*, vol. 5, issue 6, (Part -3), pp. 86-91, June 2015.

[9]. P. Chheda, M. Faruqui, and P. Mitra, "Handling OOV Words in Indian-language – English CLIR," in Advances in Information Retrieval, R. Baeza-Yates et al. (Eds.), Lecture Notes in Computer Science, vol. 7224, Springer, Berlin, Heidelberg, pp. 510-513. DOI 10.1007/978-3-642-28997-2_45, 2012.

[10]. T. Birhane and B. Hailu, "Design and Implementation of IR System for Tigrigna Textual Documents," International Journal of Modern Education and Computer Science (IJMECS), vol. 11, no. 11, pp. 31-38, 2019. DOI: 10.5815/ijmecs.2019.11.0.