iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IX **Month of publication:** September 2023

DOI: https://doi.org/10.22214/ijraset.2023.55879

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Kannada to English Agricultural Cross-Lingual Retrieval: Enhancing Knowledge Access in Farming Practices

Divija L[1], C G Hanisha Reddy[2], Rayudu Srishti[3], Dr. Surabhi Narayan[4]
[1, 2, 3]*Computer Science & Engineering, PES University, Bangalore, India*
[4]*Professor, Computer Science & Engineering, PES University, Bangalore, India*

Abstract: ***This paper presents the development and implementation of a specialized cross-lingual information retrieval system tailored for agriculture-related queries in the Kannada language. The primary objective of the system is to facilitate accurate translation of Kannada queries into English, the target language, and to retrieve relevant documents containing vital agricultural information. The proposed system addresses key challenges including incorporating effective query preprocessing techniques, designing an efficient document retrieval mechanism, and establishing optimal data indexing strategies as well as strategies have been introduced to mitigate the challenges posed by the diversity of Kannada dialects. By overcoming the language barrier, this system enables seamless and effective knowledge dissemination in the agriculture domain.***
Keywords: Cross-Lingual Information Retrieval, Agriculture, Kannada, Document Retrieval, Query Preprocessing, Data Indexing, IndicTrans.

I. INTRODUCTION

The field of Information Retrieval (IR) has emerged as a crucial discipline within computer science, dedicated to developing efficient and effective techniques for extracting information from vast datasets. The primary objective of IR is to promptly provide users with the most pertinent documents, web pages, or pieces of information in response to their queries, facilitating access to the knowledge they seek. A foundational technique in Information Retrieval is the utilization of inverted indexing, a robust data structure enabling efficient and speedy search operations. Unlike traditional forward indexing, which associates each document with its words, inverted indexing adopts a different approach. It organizes terms as keys, with each key linked to a list of document identifiers (e.g., filenames) containing the corresponding term. This inversion of the index structure allows for swift access to relevant documents associated with each query term, significantly reducing search time and computational complexity.

In the middle of the 1970s, the research of information retrieval methods first emerged. However, Salton didn't conduct a ground-breaking inquiry in this area until 1973, which was a big milestone of 46 years. Cross-language information retrieval (CLIR), which is now one of the most important research fields in the field of information retrieval, underwent significant breakthroughs particularly around 1990. The fact that CLIR is a thriving area of study has drawn a lot of interest, resulting in the publication of several investigations and studies. French, Arabic, German, Spanish, Chinese, and Italian are all included in TREC, whereas Swedish, French, Dutch, Spanish, Italian, Russian, and German are all included in CLEF. Utilizing the query translation method and result interpretation might be an efficient technique to tackle the problem of language barriers in Cross Language Information Retrieval (CLIR). Amidst the diverse applications of IR, specialized Cross-Lingual Information Retrieval (CLIR) systems have garnered substantial attention [6]. These systems aim to overcome language barriers and facilitate seamless access to information across linguistically diverse user communities. In the context of agriculture-related queries, such a specialized CLIR system holds immense potential to bridge linguistic gaps and enable effective knowledge dissemination in the agriculture domain.

This paper presents the design and implementation of a novel specialized CLIR system tailored explicitly for agriculture-related queries in the Kannada language. Kannada, a Dravidian language predominantly spoken in the Indian state of Karnataka, bears significant importance in the agricultural context due to the region's agrarian nature. The system's core objective is to accurately translate Kannada queries into English, the target language, and efficiently retrieve pertinent agricultural documents, thereby empowering users with access to relevant information irrespective of their linguistic background. The Figure 1 shows if Kannada the query is given and the related data information described in English.

To achieve this, the proposed system integrates advanced machine translation techniques to ensure accurate and contextually appropriate translations from Kannada to English. Additionally, the system incorporates query preprocessing techniques to enhance the quality and relevancy of translated queries. Furthermore, leveraging the power of inverted indexing, the system adopts an efficient document retrieval mechanism and data indexing strategies to expedite the retrieval process and streamline access to valuable agricultural insights.

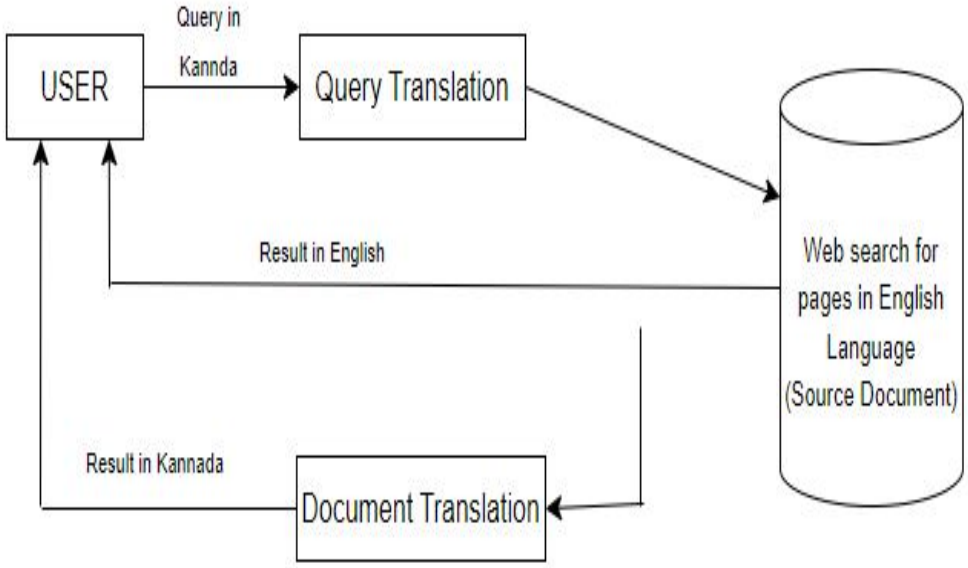


Figure 1: Cross Language Information Retrieval System

II. RELATED WORK

Mallamma V Reddy et.al [1] presented Kannada-English and Telugu-English CLIR systems, focusing on the retrieval of pertinent information for questions presented in native languages. When the query terms are not found in the dictionary, they use rule-based transliteration in addition to query translation utilizing bilingual dictionaries. A term-term co-occurrence statistic-based iterative page-rank technique is used to disambiguate numerous translation possibilities. The authors test translated queries against native document collections, documenting performance gains and analyzing the outcomes. By offering insights on how to successfully cross language barriers in information retrieval settings, the research makes a contribution to the area.

The paper Kalyani Lokhande et.al[2], discussed the development of an English to Marathi Cross Language Information Retrieval (CLIR) system, which enables users to retrieve Marathi documents using English queries. The proposed system employs query translation, matching terms in the query with Marathi documents. To enhance performance, the system incorporates query pre-processing and query expansion using WordNet. With the increasing availability and usage of multilingual content on the World Wide Web, this research addresses the need for CLIR systems for Indian languages, focusing on English-Marathi retrieval.

As mentioned by P.Iswarya et al [3], explored machine translation and ontological tree in the context of Tamil to English translation. They conducted experiments on 200 documents related to the "festival" domain. The researchers successfully developed a generic platform for bilingual Information Retrieval (IR), which showed potential for extension to any foreign or Indian language while maintaining high efficiency. This platform holds promise in overcoming language barriers and facilitating effective knowledge access and sharing across linguistic boundaries.

III. CHALLENGES IN CROSS-LANGUAGE INFORMATION RETRIEVAL

A. Ambiguity

Ambiguity develops as a result of language's inherent complexity, where a single word or phrase may have several meanings or interpretations depending on the context.

B. Phrase Identification and Translation

It might be difficult to recognize phrases in constrained contexts and interpret them based on their general meaning rather than how each individual word is used.

C. Translation Model

There is no hundred percent translation model which makes it difficult for accurate query translations.

D. Out-of-Vocabulary (OOV) Problems

Unfamiliar words get added to language which may not be perceived by the framework.

IV. IMPLEMENTATION

The implementation of this project involves developing a specialized cross-lingual information retrieval system tailored for agriculture-related queries in the Kannada language. To achieve this, the system addresses key challenges, including effective query preprocessing, an efficient document retrieval mechanism, and optimal data indexing strategies.

A. Text Preprocessing

The crucial step in the implementation is text preprocessing. Before creating the inverted index, the text documents need to be preprocessed to ensure uniformity and to eliminate noise that might affect retrieval accuracy. The preprocessing steps include:

1) *Lowercasing*: All the text is converted to lowercase to ensure case-insensitivity during the retrieval process. This way, queries in any letter case can be matched effectively.
2) *Removal of Non-Alphabetic Characters*: Non-alphabetic characters and special symbols are removed from the text. This step helps eliminate punctuation and other symbols that do not contribute to the retrieval process.
3) *Tokenization*: The preprocessed text is tokenized into individual words or terms. Tokenization is crucial as it breaks down the text into meaningful units, making it easier to create the inverted index.
4) *Stopword Removal*: Common words that do not carry significant meaning, such as "and," "the," "is," etc., are removed from the tokenized text. Stopword removal reduces the size of the index and focuses on more meaningful terms.

B. Inverted Indexing

The heart of the Information Retrieval System is the inverted indexing technique. Once the text preprocessing is completed, the implementation proceeds to construct the inverted index. The inverted index is a data structure that maps each unique term (word) in the document collection to the list of document identifiers (e.g., filenames) that contain that term. For each term in the tokenized and preprocessed text, the implementation records the document identifier(s) in which the term occurs. The inverted index enables fast retrieval of documents associated with a particular query term, as it avoids the need to scan every document for matches.

C. Translation of Kannada Query to English

As part of the implementation, the system incorporates a translation module that can translate Kannada queries entered by users into English. This step is essential for ensuring language compatibility and enabling efficient retrieval in English-based document collections. By translating the query to English, the system can effectively match the query terms with the terms in the inverted index, which is constructed using English text. This translation is done using IndicTrans.At the time of writing (14 April 2021), Samanantar dataset, the biggest publicly accessible parallel corpus collection for Indic languages, was used to train the multilingual Transformer-4x NMT model known as IndicTrans.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

D. Query Processing

After translating the Kannada query to English and preprocessing it, the system processes the query to identify the relevant terms. The query is then tokenized and preprocessed in the same way as the text documents. The implementation then looks up the inverted index for each query term to obtain the list of document identifiers containing those terms.

E. Query Expansion

Query expansion is a crucial technique in Information Retrieval that aims to improve the search performance by broadening the scope of the initial user query.

The process involves augmenting the original query with additional related words, synonyms, or conceptually similar terms obtained from lexical resources such as thesauri or WordNet. By incorporating these expanded terms into the query, the system aims to retrieve more comprehensive and relevant search results, increasing the chances of finding documents that might not have been captured by the initial query.

Query expansion is particularly useful in overcoming issues like polysemy and synonymy, where a single word can have multiple meanings or different words can represent the same concept. Through the integration of additional terms, the query becomes more robust and contextually enriched, contributing to the enhancement of search accuracy and user satisfaction in information retrieval systems.

F. Ranking of Document:

After obtaining the relevant documents for each query term, the retrieved documents are ranked based on their relevance to the user's query.

This ranking is typically done using a scoring mechanism such as TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF assigns a score to each document based on the frequency of query terms in the document and the rarity of the terms in the entire document collection. The implementation ranks the documents based on their TF-IDF scores, presenting the most relevant documents at the top of the search results.

The implementation of the Information Retrieval System combines text preprocessing, query translation, inverted indexing, query processing, and document ranking to efficiently retrieve relevant documents in response to agriculture-related queries in the Kannada language.

By addressing language compatibility and utilizing powerful indexing and retrieval techniques, the system provides users with quick and accurate access to the knowledge they seek in large datasets.

As mentioned earlier, ambiguity is one key challenge faced in a CLIR. One approach to solve this problem of ambiguities is by building a new dialect dataset that mainly focuses on different agricultural terms collected from native speakers of Kannada language.

Then the ambiguous word can be identified and replaced with official Kannada word which makes the translation seamless.This approach has been discussed by N.B.Chittaragai et.al [4].

V. RESULTS

Figure 5.a : For the query given in Kannada
Figure 5.b : Translated Kannada query to English
Figure 5.c : Tokenized terms in English Query
Figure 5.d : Top K- Relevant Documents for the given query
Figure 5.e : Choosing option 1 will preview of the document
Figure 5.f : Choosing option 1 will display of the entire document

Fig. 5 a. Query given is “ತೆಂಗಿನಕಾಯಿಗೆ ಯಾವ ಕೀಟನಾಶಕಗಳನ್ನು ಬಳಸಲಾಗುತ್ತದೆ?”

```
query = input("Please enter your query in kannada: ")
q=[query]
print("translated query(kn -> en):")
knq=indic2en_model.batch_translate(q, 'kn', 'en')
```

```
Please enter your query in kannada: ತೆಂಗಿನಕಾಯಿಗೆ ಯಾವ ಕೀಟನಾಶಕಗಳನ್ನು ಬಳಸ
translated query(kn -> en):
100%|██████████| 1/1 [00:00<00:00, 1351.26it/s]
```

Fig. 5 b. Translated Kannada query to English

```
knq
```

```
['Where can we grow coconuts in India']
```

Fig. 5 c. Tokenized terms in English Query

```
[nltk_data] Downloading package wordnet to /root/nltk_dat
[nltk_data]   Package wordnet is already up-to-date!
Words used in information retrieval:
['Where', 'can', 'we', 'grow', 'coconuts', 'in', 'India?']
```

Fig. 5 d. Top K- Relevant Documents for the given query

```
Relevant Documents (in decreasing order of relevance):
1. 'Coconut.txt'
2. 'Planter_(farm_implement).txt'
3. 'grow-coconut-palms-inside-1902595.txt'
4. 'golden-chain-trees-2132131.txt'
5. 'Manure_spreader.txt'
6. 'Gravity_wagon.txt'
7. 'Threshing_board.txt'
8. 'Monowheel_tractor.txt'
9. 'Insecticide.txt'
10. 'Echeveria.txt'
================================================================================
```

Fig. 5 e. Choosing option 1 will preview of the document

```
Coconut.txt:
1. Display a preview
2. Read the full document
Document ID: Coconut.txt


The coconut tree (Cocos nucifera) is a member of the palm tree family (Arecaceae) and the only living s

The coconut tree provides food, fuel, cosmetics, folk medicine and building materials, among many other
```

Fig. 5 f. Choosing option 1 will display of the entire document

```
================================================================================
================================================================================
Coconut.txt:
1. Display a preview
2. Read the full document
Full Document ID: Coconut.txt


The coconut tree (Cocos nucifera) is a member of the palm tree family (Arecaceae) and the only livi

The coconut tree provides food, fuel, cosmetics, folk medicine and building materials, among many o

The coconut has cultural and religious significance in certain societies, particularly in the Austr

Coconuts were first domesticated by the Austronesian peoples in Island Southeast Asia and were spre

Trees grow up to 30 metres (100 feet) tall and can yield up to 75 fruits per year, though fewer tha

Cocos nucifera is a large palm, growing up to 30 metres (100 feet) tall, with pinnate leaves 4–6 m

True-to-type dwarf varieties of Pacific coconuts have been cultivated by the Austronesian peoples s

Botanically, the coconut fruit is a drupe, not a true nut.[13] Like other fruits, it has three laye

The interior of the endocarp is hollow and is lined with a thin brown seed coat around 0.2 mm (1/64
```

VI. CONCLUSION

In summary, this paper introduces a dedicated CLIR tailored for agriculture-related queries in Kannada. By effectively translating these queries into English and utilizing foundational Information Retrieval techniques like inverted indexing, the system overcomes language barriers to provide seamless access to vital agricultural information. Through the integration of advanced machine translation, query preprocessing, and efficient document retrieval mechanisms, this system contributes to bridging linguistic gaps and empowering users with valuable insights irrespective of their language proficiency. Looking ahead, the outlined future directions encompass a comprehensive approach to further enhance the system's effectiveness. Incorporating regional dialects, implementing user feedback,etc mechanisms will collectively refine the translation process. Additionally, extending multi-Indian language support, integrating real-time agricultural data sources, and prioritizing user-centric interface design will ensure continued relevance and usability in the dynamic landscape of agricultural knowledge dissemination.

VII. FUTURE WORKS

A. *Inclusion of Kannada Dialects*

Incorporate different Kannada dialects, like Hubli, Mysore, and Mangalore, into the translation system to cater to users from diverse Kannada-speaking regions.

B. *User Feedback Mechanism*

Implement a user feedback mechanism to enable users to rate the quality of translations and provide feedback on specific phrases. Use this feedback to fine-tune the translation model and improve the system's accuracy iteratively.

C. *Robustness and Adaptability*

Enhance the translation system's robustness by incorporating advanced neural network architectures, such as transformer-based models, and investigate methods like self-supervised learning for better adaptation to domain-specific vocabulary.

D. *Multi-Indian Language Support*

Extend the translation system to support more Indian languages, such as Tamil, Hindi, Bengali, and Telugu, to facilitate communication among speakers of various languages in the agriculture domain.

E. *Expanding the Dataset*

Expand the dataset by collecting, curating more agricultural-specific documents in Kannada and English, including research papers, crop advisories, farming manuals, and market reports, to improve the translation system's domain expertise.

F. *Real-time Agricultural Data Integration*

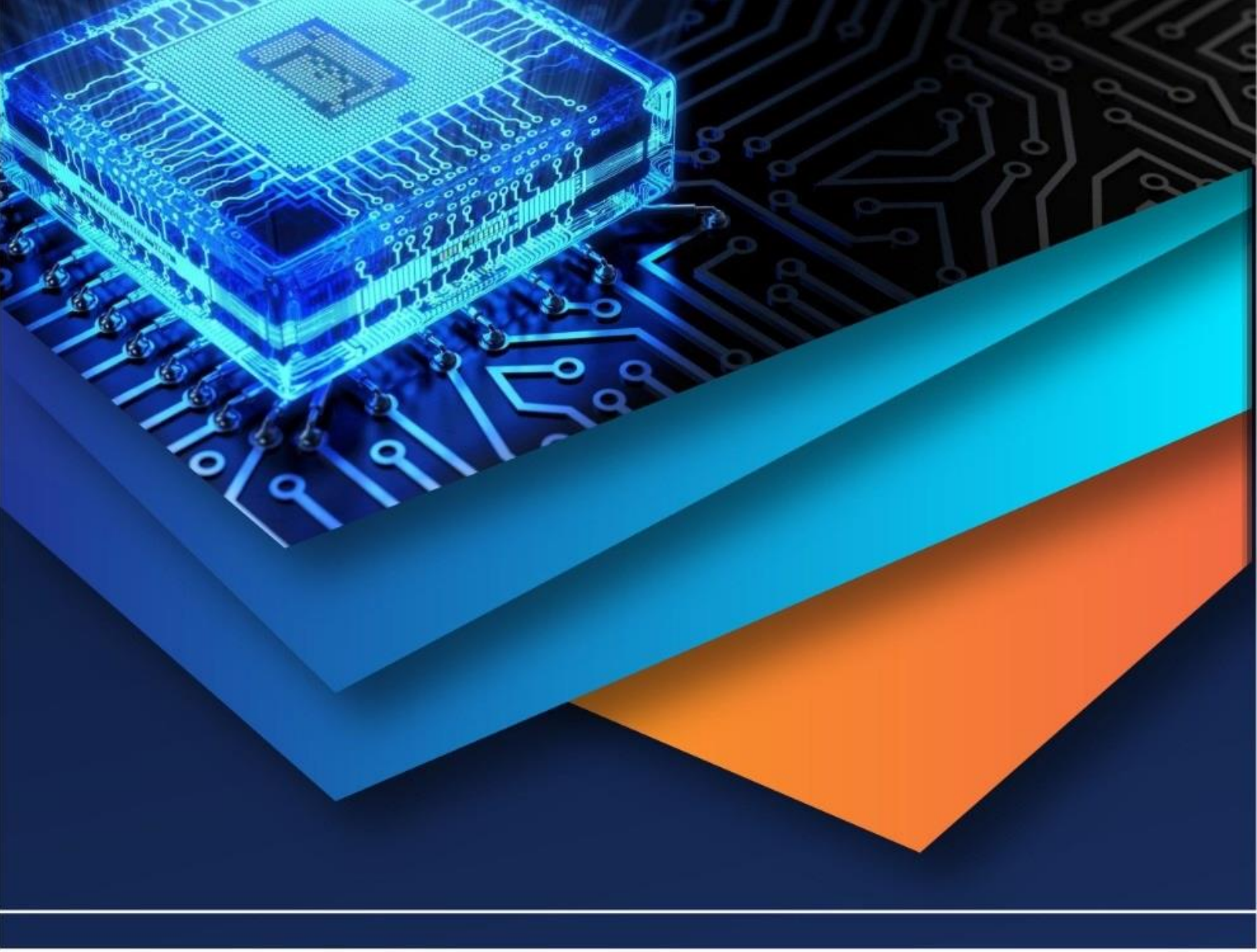
Investigate methods to integrate real-time agricultural data sources, such as weather APIs, crop yield predictions, and market prices, into the translation system, ensuring users receive timely and relevant information.

G. *User-Centric Interface*

Design an intuitive and user-friendly interface that allows farmers and stakeholders to easily access the translation system, provide feedback, and access relevant agricultural information.

REFERENCES

[1] Mallamma V Reddy et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, 1876-1880
[2] K. Lokhande and D. Tayade, "English-Marathi Cross Language Information Retrieval System," in International Journals of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 8, August 2017, ISSN: 2277-128X, pp. [112-116], DOI: 10.23956/ijarcsse/V7I8/0127.
[3] P. Iswarya and V. Radha, "Speech and Text Query based Tamil – English Cross Language Information Retrieval System," in 2014 International Conference on Computer Communication and Informatics (ICCCI -2014), Coimbatore, India, Jan. 03 – 05, 2014.
[4] N. B. Chittaragi and S. Koolagudi, "Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms," Language Resources and Evaluation, vol. 54, Jun. 2020.DOI: 10.1007/s10579-019-09481-5.
[5] Bajpai, P., & Verma, P. (2014). Cross Language Information Retrieval: In Indian Language Perspective. International Journal of Research in Engineering and Technology, 3(Special Issue 10), 46.
[6] P. Galuščáková, D. W. Oard, and S. Nair, "Cross-language Information Retrieval," arXiv preprint arXiv:2111.05988, Nov. 10, 2021, revised version: Jun. 8, 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY