Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 12, No. 2, June 2024, pp. 319~332
ISSN: 2089-3272, DOI: 10.52549/ijeei.v12i2.5172

A Translation Framework for Cross Language Information Retrieval in Tamil and Malayalam

Sakthi Vel S.[1], Dr.Priya R[2]
[1]Research Scholar, Centre for Development of Imaging Technology (C-DIT)
University of Kerala, Kerala, India
[2]Assistant Professor, Department of Computer Science, Government College Kariyavattom, Kerala, India

Article Info

Article history:

Received Oct 18, 2023
Revised May 5, 2024
Accepted May 24, 2024

Keywords:

Cross Language IR
LSTM encoder-decoder
Machine Translation
Text Processing
Query expansion
BLEU Score

ABSTRACT

Cross Language Information Retrieval (CLIR) stands as an essential element in multilingual information accessibility, enabling users to obtain relevant information even when the query language and the language of the documents diverge. This paper proposes a translation framework for CLIR in Tamil and Malayalam, two Dravidian languages widely spoken in South India. Different challenges prevail in CLIR of these languages due to their linguistic differences, translation equivalence, mapping source to target languages, semantic equivalence, limited dataset and tools for ongoing research in this domain. The proposed methodology resolves some of the issues around training of a corpus utilizing a Long Short-Term Memory (LSTM) based encoder-decoder translation model. The study incorporates two bilingual parallel corpora comprising 373 sentences pairs each. Evaluation of the model's accuracy is conducted by equivalency its translations against reference translations using the Bilingual Evaluation Understudy (BLEU Score). Furthermore, BLEU scores obtained from proposed LSTM-based encoder-decoder model is compared with those from Google Translate. The findings reveal that the LSTM model attains an average BLEU score of 0.933, where, performance of Google Translate, achieved a score of 0.813. Finally, the study conducts a comparative analysis with selected CLIR models in different languages, to evaluate the overall performance of the proposed approach.

Corresponding Author:

SAKTHI VEL S.
Research Scholar
Centre for Development of Imaging Technology (C-DIT)
University of Kerala,Trivandrum, Kerala, India
Email: sakthivels@keralauniversity.ac.in

1. INTRODUCTION

Cross Language Information Retrieval is a sub area of Information Retrieval (IR) System, in which the query language and language of the documents retrieved are different. CLIR is an interdisciplinary field of Information Retrieval, Natural Language Processing, Machine Translation, Languages and Text Processing. CLIR finds a different application which includes facilitating users to search the information without any limitation of language barriers, multilingual information access, increasing the amount of online information available in non-English languages, helping the multilingual speakers interact and collect a greater number of documents from different languages. CLIR is also useful for multilingual population regions that share multilingual documents.

CLIR systems have been divided into Bi-lingual, Multilingual and domain oriented based on different languages of query and documents. The general architecture of CLIR system can be classified into Query translation, Document translation and combination of query with documents as shown in Figure 1.

The process of CLIR system modelling consists of different approaches such as dictionary, corpus, machine learning, machine translation, deep learning, linguistic, and rule-based system. The basic techniques of CLIR in Query translation can be further classified as Dictionary, Corpus and Machine Translation based [1]. Final layer of this figure 1 again classified into two language corpora are based on collected data set.

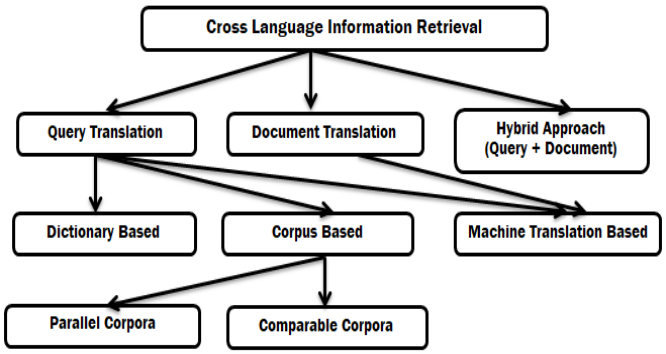

Figure 1. Techniques for Cross Language Information Retrieval (CLIR)

Dictionary based query translation involves processing the given user's query linguistically to find the relevant keywords with the help of Machine Readable Dictionary (MRD). MRD's are collections of electronic versions of either general or domain oriented printed dictionaries. In dictionary based approach of CLIR, translation is an easy task but it has lots of limitations such as ambiguities, quantity, and quality of the dictionary and many more. The corpus based CLIR techniques involve a word-by-word analysis and translation of the corpus which produces a set of translation probabilities for each term in a query. The works are centred on data collected from multilingual Corpora using various corpus and linguistics techniques [2]. It uses either parallel or comparable corpora, parallel Corpora consist of exactly same documents but in different languages. A parallel corpus is a collection corpus that contains a collection of original texts in source language (L1) and their translations into the respective target languages (L2, L3, …., Ln). A comparable corpus consists of a set of documents presented in different languages but documents are not the translations of each other.

The aim of CLIR based Machine Translation system is to translate queries and documents from one language to another by using a context [3]. Document translation is performed by any one of the Machine translations tools like Google Translate, Bing and SYSTRAN. Machine Translation systems translate both queries and documents but it requires large amounts of parallel corpora [4]. There are four different techniques to deal with Machine Translation based CLIR; such as Word-for-Word, Syntactic transfer, Semantic transfer, and Inter-lingua techniques [5], [6]. Many factors affect Machine Translation based CLIR, including Polysemy (Words with multiple meanings), sentence alignment with one-to-many relationships [7], mapping of bi-lingual dictionary, limitations of dataset, default sentence structures, different grammatical structures, inflected and derivational feature of lexicon. Machine Translation based CLIR provides a platform to retrieve foreign language documents through user's query in native language [8].

The recent research in CLIR is in ontology-based models. Ontology is a formal, specification of conceptualization, captures the structure of the domain [9], explicit specification of a common and shared conceptualization. It consists of a set of distinct and identified concepts interconnected with a set of relations [10]. Semantic based CLIR system can include different available concepts and ideas which are expressed by the user through her/his query and can thus provide more accurate results than the traditional keywords based search [11]. For example, a simple ontology for person consists of a set of concepts C_{Person}= {men, women, child, parents} and a set of relationship R_{Person}= {Father, Mother and Son} [12]. Unlike the other two models, ontology could attain minimal error performance and maximum precision.

Though, there exist many approaches, still there are gaps to attain precision in full implantation of the required query of users. The entitled research paper attempts to identify the limitations in the existing CLIR models such as translation equivalence, ambiguity, and grammatical differences, lexical and semantic variations in Tamil and Malayalam and implemented LSTM based translation framework for better precision. This article is organized into five sections, where section 2 describes related works in detailed manner. Then section 3 proposed a model for CLIR system, and section 4 discussed results with related discussions. Finally, section 5 includes the conclusion of this article.

2. RELATED WORKS

Kumar et al. [13] authors reviewed some of the recent works related to CLIR system and the study proposed architecture of English-Hindi CLIR system. The system implemented cross-lingual web querying using bi-lingual ontology's of English to Hindi. Zeeshan et al. [14] proposed Neural Machine Translation (NMT) based Chinese to Urdu (C2U) word base dictionary machine translation. They designed electronic dictionary containing 24,000 entries from Chinese to Urdu. The corpus has been trained by two NMT models such as LSTM and Transformer. Finally, the study concluded LSTM gave 0.06 to 0.41 and Transformer gave 0.07 to 0.52 BLEU score.

Aditi et al. [15] studied different metrics used in Hindi machine translation. They listed several machine translation evaluation metrics such as BLEU, Rouge, METEOR, TER, and METEOR. Nikesh et al. [16] worked for English-Malayalam CLIR system. The system involves retrieval of Malayalam documents through an English query. The system processed English queries, using University of Massachusetts (UMass's) stop word list and stemmer algorithm. For ranking of retrieval documents, Vector Space Model (VSM) was used and to test the system 25 queries were used.

Ibrahim et al. [17] proposed language modelling based Amharic-Arabic CLIR system that accepts a text query in Amharic from the user, translates it into Arabic language using the, pre-trained Neural Machine Translation model and then searches for both Amharic and Arabic language documents using language modelling based retrieval model and enhance the retrieval performance by incorporating Parts-of-Speech Tagging model as the optimization techniques of the ranking algorithms. Finally, the paper concluded with Mean Average Precision (MAP) of Amharic as 0.8833 and that for Arabic as 0.93.

An ontology-based Tamil-English cross lingual information retrieval system using Word Sense Disambiguation (WSD) to resolve the ambiguity of Tamil query to English was implemented [18]. In the work, the root word and its corresponding suffix marker were identified. Later multiple meaning and ambiguities were resolved through manually constructed WSD, which was followed by rearrangement of syntactic structure and finally query reformation of the targeted documents. A Tamil-English bilingual dictionary of size 6.08 MB related to agriculture domain was used in the study. A precision of 95.36% was obtained for top 20 pages retrieved by the Google search engine.

The author proposed Multilingual Information search for three languages: English, Hindi and Malayalam using a Novel approach [19]. The study consists of five major tasks of CLIR such as Query pre-processing, Searching, Processing webpage contents, retrieval, and ranking. Subtasks of query pre-processing such as language detection, stop word elimination, and stemming. After pre-processing, the query words are passed on to searching module. Finally, relevant Web Pages contents are collected, followed by assigning ranks to each retrieved document. In this experiment, total of 30 queries are submitted to the system for evaluating the performance and an average precision of 0.539 was obtained.

An experiment for two languages in Tamil and English through Conceptual based search engine system was performed [20]. The model consists of two-layer architecture namely online and offline process. The system mainly focused on dictionary based approach and proposed a model involving pre-processing, query expansion, translation from English to Tamil, concept-based indexing, searching, and ranking. Another work proposed a system that tested for tourism domain with 50,690 documents corpus (25,690 Tamil documents and 25,000 English documents) and achieved 0.51 precision for both Tamil and English queries (20 queries each for Tamil and English).

A novel approach was proposed for an improved English-Hindi CLIR system [21]. The system employed Naïve Bayes and Particle Swarm optimization for an efficient CLIR system of the given languages pair. In another work in the same year, [22], an unsupervised corpus based WSD of Marathi-English CLIR system was proposed. The dataset has been collected from 2011 Forum of Information Retrieval Evaluation (FIRE). The system consists of four important components; pre-processing of query, query translation, transliteration and WSD. The system achieved 0.73 average recall values from 15 average numbers of relevant documents.

In 2010, architecture was proposed for bilingual information retrieval system for English and Tamil. The study was mainly divided into four different modules, such as User interface, Keywords extraction, Information retrieval and extraction and output display [23]. The ontological tree model is used to identify and match each keyword in the given language. Authors collected and developed ontological tree in "Festival" domain and more than 200 documents were collected from these two languages. The model was evaluated using precision, recall and F-measure values. The authors finally concluded that, bilingual search engine through ontological tree was improved by 40% for English and 60% for Tamil language.

In [24], the work proposed two CLIR processing steps such as pre-processing and post-processing. Pre-processing steps consists of the sub tasks of Query Parsing, Query Expansion, Query Formulation and Search Knowledge Sources. After pre-processing the authors discussed stages of post-processing such as Parameter estimation, Categorization, Aggregation, display processing and finally learning the parameter

from the given documents. The system proposed architecture for Hindi-English CLIR using query expansion to improve the relevancy of retrieval documents. In the first experiments, Query expansion is performed with and without OkapiBM25 ranking algorithm. The results show that the relevancy in terms of Mean Average Precision of retrieved documents is higher with OkapiBM25 as compared to the one without ranking [25].

An outline model of Arabic and English multilingual ontology to improve the query translation in "Travel" domain was proposed in another work. The model created a domain of travel ontology consisting of 100 English concepts mapped to the Arabic concepts. Finally, the model evaluated and compared the MAP of both Machine Readable Dictionary and Ontology of these two languages. The paper concluded with average MAP for machine readable dictionary as 0.42 and that for ontology as 0.63 [26].

Another work focused on a cross analysis of CLIR using various approaches for Indian languages [27]. The authors list out major cross language information retrieval system in English related with Indian languages. The review details on the works carried out in English to Hindi, Tamil, Malayalam and so on. Another work introduced semantic search rather than keyword based search [28]. The system offered input query as English and output documents in Hindi or Bengali. The system designed user interface with the help of Tkinter tool. Finally, the system was employed but searched only one word "Narendra Modi".

In the study employing unsupervised Cross-Lingual Information Retrieval using monolingual data, a fully un-supervised model for ad-hoc CLIR system which requires only multilingual data was developed [29]. The above model was constructed which could map queries or documents into embedding space structures. The paper discussed three techniques related to word embedding space such as Cross-Lingual Embedding from Comparable Documents (CL-CD), Cross-Lingual Embeddings from Word translation pairs (CL-WT), and Cross-Lingual Embeddings without bilingual super vision (CL-UNSUP). They have experimented standard CLEF-CLIR dataset collected from three languages of English to Dutch/Italian/Finnish. Finally, the paper proposed and developed cross lingual embedding space methods of these languages with three different data alignment such as document-aligned comparable data, word translation pairs and no bilingual data. The system achieved Mean Average Performance for all three language pairs: English to Dutch-0.336, English to Italian-0.347 and English to Finnish- 0.307.

An evaluation method to automatically discover links between Japanese to Chinese using Japanese-Chinese Cross-Language entity linking method was proposed [30]. Here the study consists of two steps, initially author translates Japanese key phrase into Chinese documents and finally, the original Japanese documents are translated into Chinese documents. The authors evaluate cosine similarity between all original articles of these two languages. The paper used data set from Wikipedia and Baidu Baike. Baidu Baike is a large collection of more than fifteen thousand Chinese article for free of access. The paper concludes that, the system achieved the accuracy rate of 97% by using Baidu Baike and accuracy rate of 81% by using Wikipedia.

In [31], a model of E-learning multi-language ontology of English and Spanish was proposed. The aim of this model was to construct a domain specific multi lingual retrieval system. E-learning-Course/Lecturer from MIT open courseware for both English and Spanish was the domain used for study. Another work describes various methods of bi-lingual CLIR system from online documents [32].The author suggests translating search queries in a local language into English, after retrieving relevant documents in both languages. Korean to English conversion through both transliteration and back-transliteration methods are also employed in the study.

A combination of Keyword-based IR with a Latent semantic-based model for Arabic-English CLIR system using Deep Learning was proposed by [33]. The proposed CLIR system used Deep Belief Networks (DBNs) to identify the latent semantic of Arabic queries into English documents. The author experimented and evaluated with three parameters of λ, β, and K values. λ represent weight of the lexical-matching score, β- represent weight of the Arabic DBNs score, and K-represent number of top most search documents. The system achieved highest accuracy of 91.6% for combined Semantic analysis with Lexical matching.

Indonesian-Japanese [34], discussed on term extraction from Indonesian-Japanese Bi-lingual corpora using machine learning algorithm. They introduced new methods for term extraction between these two languages within three criteria such as first n-gram extraction for Indonesian and Japanese, n-gram cross pairing between these two languages and finally classification of extracted documents. The sub components of this system are corpus pre-processing, term pair extraction, feature extraction and classification of documents. The system evaluated three different features such as linguistic, statistic, and combination of both features. Finally, the system achieved 98.6% overall accuracy.

In 2010, another work was implemented which evaluated a synergistic approach between Thesaurus-based approach and Corpus-based approach [35]. A study was done on E-learning domain ontology of English and Spanish for automatic semantic mapping between these two languages. The study concluded by evaluating the concepts and sub-concepts extracted from English and Spanish. The system

performed 0.92 Average Top-50-Recall and 1.00 Average Top-50-Precision from top 50 randomly selected documents.

Table 1. Different models of CLIR systems

Language & Domain	Algorithms	Observations
Tamil-English & Agriculture	Ontological-based WSD	Mean Average Precision of 95.36% for top 20 Pages.
English-Dutch/ Italian/ Finnish & CLEF (Cross Language Evaluation Forum)	Unsupervised CLIR word embedding space	MAP of three models of both values of the interpolation factors (λ= 0.5 and λ= 0.7).
Amharic-Arabic & New Domain only	Pre-trained Neural Machine Translation with Parts-of-Speech Tagging (POS)	Mean Average Precision of Amharic is 0.8833 and Arabic is 0.93. Three POS tag set are used: CRF, Brill and TnT.
English, Hindi/ Malayalam & Word: Lakshadweep	Multilingual Information Search Algorithm- A Novel Approach	The average Precision is 0.53973790 with top 30 related queries.
Japanese-Chinese & Encyclopaedia (Chinese & Japanese)	Cross Language Entity Linking (CLEL)	Accuracy rate of 97% by using Baidu Baike and accuracy rate of 81% by using Wikipedia.
Indonesian-Japanese & Computer Science	Term extraction form Indonesian-Japanese Bi-lingual corpora using machine learning algorithm.	The system achieved 98.6% overall accuracy, 4.96% precision and 24.47 % recall values.
English-Hindi & Query: "Machine Learning Research Group"	Improved approach through n-gram model	More number of relevant documents in the first n/3 retrieved documents.
Marathi-English & FIRE 2011 (Forum of Information Retrieval Evaluation)	Detailed user query and unsupervised corpus-based WSD	0.73 average recall and 0.045 average precision values form 15 average numbers of relevant documents.
Hindi-English & Selected sentences from both language	Query translation with Ambiguity removal in both language grammatical structure	The precision of the selected query is 0.83 (Number of relevant documents is 10 out of top 12 retrieved documents appeared on first page.)
Hindi-English & collected from Forum for Information Retrieval Evaluation (FIRE)	Query expansion with Word Sense Disambiguation and without OkapiBM25 ranking method	MAP of top 10 queries before query expansion is 0.5379 and after query expansion is 0.6742.
English-Hindi/ Bengali & Word: Narendra Modi	Query translation using semantic search	System improved quality and relevancy of the search result.
Tamil-English & Tourism	Dictionary with Concept based search engine	0.51 of precision for both Tamil and English queries (20 queries for each Tamil and English).
English-Hindi & Bi-lingual ontology	Cross-lingual web querying using bi-lingual ontology	Average precision of top 100 documents is 0.1064.
Arabic-English & Selected articles from Wikipedia	Deep Belief Networks with latent semantic-based CLIR	Experimented and evaluated with three parameters of λ, β, and K values.
English-Spanish & Ontology: E-learning-(Course/Lecture)	Synergistic approach between Thesaurus & Corpus based approach	Corpus-based approach achieved highest accuracy.
English-Spanish & E-Learning (Ontology)	Synergistic approach between Thesaurus & Corpus based approach	The Average of Top-n-Recall is 0.92 and average Top-n-Precision is 1.00.
English-Tamil & Festivals	Ontological Tree	40% for English and 60% for Tamil language.
Arabic-English & Travel	Multilingual Ontology as Translation of Query	Mean average precision of both MRD and Ontology is 0.7.
English-Persian & Ontology of Bi-lingual Dictionary	Hybrid approach of Bi-lingual query translation	Three levels: word (precision: 0.4), Phrases-based translation (precision: 0.5) and improved-based translation (precision: 0.6).
Korean- English & Word: Samsung	Combined with transliteration and back-transliteration methods	Average precision of these two transliterations is 0.6.
French- English & Canadian Parliament	Query-translation from parallel texts	Around 80-90% of the retrieval effectiveness.

Another work proposed a model for CLIR on English-Persian with focus on solving ambiguity [36]. The research work attempts to solve the issue over ambiguity by proposing a combination of syntactic and semantic model to improve the dictionary-based translation. The task of query processing of Persian into English consists of three different levels such as, the selection of word translation, query expansion and finally comparison and evaluation of the system. The performance of the system was evaluated on three levels; Word-based translation (precision: 0.4), Phrases-based translation (precision: 0.5) and Improved-based translation (precision: 0.6).

Here the aim is to bring out the available relevant models of CLIR and its details about their techniques and performance accuracy. In table 1, includes some cited available CLIR works. The research works carried out in various language pairs, domains, algorithms and accuracy of each work are given in below.

Different techniques had been used by researchers for obtaining high performance in CLIR systems. The researcher had come across the CLIR methods such as Dictionary based, Corpus, Ontology, translation, linguistics and Machine learning based while reviewing the above listed works. Based on the studies done, the following findings could be recognized at different approaches.

- The issues identified with dictionary based CLIR are: Out-Of-Vocabulary (OOV) words, meaning of the words that are not found in the dictionary, lexical ambiguity in the source and target languages.
- Limited collection of lexicons, and inflectional issues of the concerned pair of languages. The limited collection of lexicon implies that the possible number of lexicons in dictionary may not be sufficient to satisfy the amount of user's language queries.
- The availability of keywords may disturb the accuracy of translation in language pair and it may affect the performance of the source and target languages.
- The daunting task in dictionary-based model is dealing with identification of grammatical complexities, inflection, pre-fixation, suffixation which vary according to the languages. Therefore, the highly inflectional languages like Tamil and Malayalam pose greater challenges.
- Ambiguities lead to inaccurate performance of the model and thereby end in poor results.
- Some other minor issues such as spelling variants, homonyms, and hyponymy continue to limit the performance of dictionary based CLIR system in the above cited works.
- The issue over the lexical ambiguity occurs due to the available size of dictionary. This could be avoided only through ample size of the lexicon set, semantic data, data size, labelling of data etc.
- To provide the minimal error result, the model must be of enough size and grammatically tagged.
- Most of the corpuses are domain oriented not generalized.
- The short comings are not only limited to the size and qualities of the corpus, but the aspects internal to the corpus such as unstructured pattern, ambiguity, number of occurrences etc.
- The major issues of corpus based CLIR, is due to segmentation of text into sentences.
- Corpus updating and modifying data items are also a complex task in this approach.
- To build ontology, each time we must construct separate decision tree. So, it will be more complex relationship and time consuming.
- It is a complex data structure mapping from source language to target language.
- The rules are not enough to transform relational database to ontology.
- The mapping process with real-world objects into corresponding meaning is a challenging task.

3. RESEARCH METHOD

Most of the proposed techniques used for cross searching among the languages are based on rule-based model combined with linguistics principles. This article, proposed a model for cross language information retrieval using translation-based modelling for two specific languages such as Tamil and Malayalam. Figure 2 shows the model with different components such as dataset creation, Query expansion, Query translation, Parts-of-Speech tagging of query and documents of source and target languages, Language Modelling of Query, and documents, CLIR searching and Ranking modules and Final documents evaluations. The three modules of the architecture, corpus collection, query expansion and query translation, are trained and evaluated in this work.

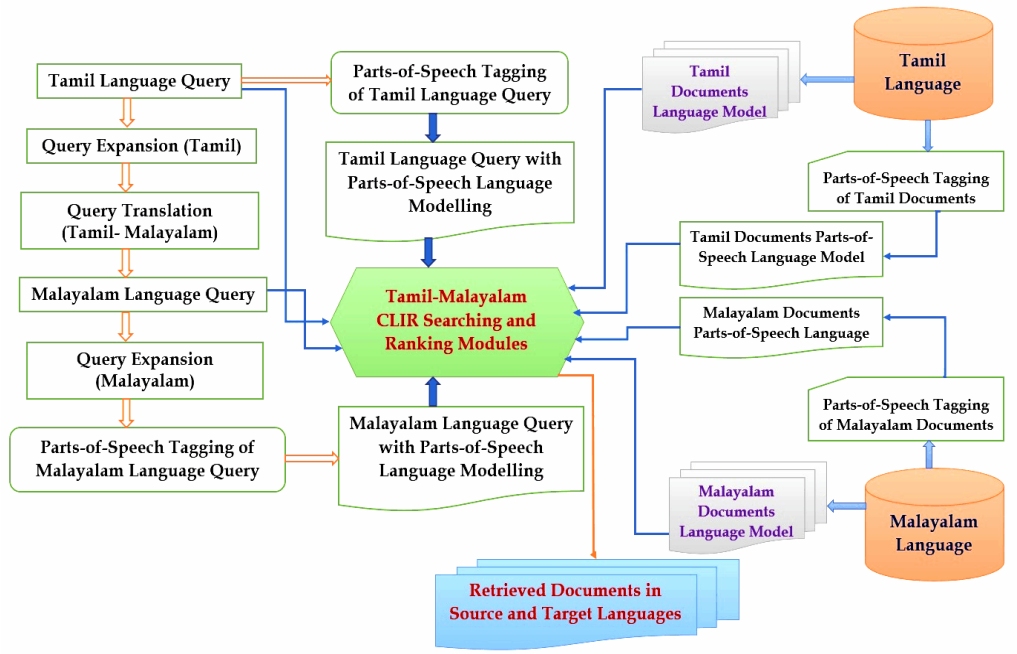

Figure 2. Proposed model for CLIR Tamil and Malayalam system

3.1. Corpus Collection

A small set of parallel text corpora are utilized in this model. The two types of online datasets used for the experimentations process are manually collected from CLARIN-ERIC (European Research Infrastructure Consortium) which provides an online repository of parallel corpora for various languages for training data for statistical machine translation process [37], [38]. The following table shows the total number of sentences from each document and its translation equivalence between source and target languages. Here, English-Tamil bi-lingual consist of three types of sentences based on simple sentences, sports, and person to person interactions such as type 1, type 2 and type 3. The second corpora Malayalam-Tamil bi-lingual consist of two types of sentences based on television news and daily conversation such as type 1 and type 2.

Table 2. Total Number of Sentences from collected corpora

Languages	Types of Sentences	Total Number of sentences
English-Tamil bi-lingual Parallel Corpora [37]	Type 1: 100 Type 2: 82 Type 3: 19	**201**
Malayalam-Tamil bi-lingual Parallel Corpora [38]	Type 1: 50 Type 2: 122	**172**
Total sentences from both Corpora		**373**

3.2. Query Expansion

Initially, the raw text is transformed into a suitable form for higher-level processing. Some types of techniques are required to expand and pre-process the above given documents. The steps for query expansion used are:

- Tokenization into word and sub-word level
- Stop Word Removal (Commonly and frequently used words)
- Stemming (Finding and removing all possible suffixes)
- Add start and end tokens of target language, and
- Create a vocabulary of source and target languages

Algorithm 1. Shows that overall procedures used in the text pre-processing of the collected documents [39].

Algorithm 1: Query expansion and Pre-Processing

Begin
Input: Array of two Corpora (English-Tamil and Malayalam-Tamil)
Output: Array of pre-processed two Corpora
Procedure: Steps in Pre-Processing
For each Corpora in document, Do:
Load the Documents
Tokenization of Sentences and Words
Remove Stop Words
Stemming
Add start and end tokens to target language
Create a vocabulary of source and target languages
Update Document to New Corpora
Return all Documents
End Procedure
End Until
End

After query expansion, the source and target vocabulary created are applied into model training and evaluation. Figure 3 and Figure 4 show, results after pre-processing of both documents. The processed documents consist of a reference id, sentence type, maximum length, start and end marker of target for each sentence.

1 to 10 of 10 entries Filter

index	source	english_sentence	tamil_sentence	length_eng_sentence	length_tam_sentence
77	Type2	I have to leave now.	START_ நான் இப்பொழுது கிளம்ப வேண்டும் _END	5	6
30	Type2	I want to sleep.	START_ நான் தூங்க விரும்புகிறேன் _END	4	5
197	Type2	People who live in glass houses shouldn't throw stones.	START_ கண்ணாடி வீட்டில் வசிப்பவகள் கல்லை எறியக் கூடாது _END	9	8
73	Type2	Don't listen to her.	START_ அவள் சொல்வதைக் கேட்காதீர் _END	4	5
139	Type2	Is he a friend of yours?	START_ அவர் உங்களுடைய நண்பரா _END	6	5
81	Type2	I'm taller than you.	START_ நான் உன்னை விட உயரமாக இருக்கிறேன் _END	4	7
129	Type2	We ran after the thief.	START_ நாங்கள் திருடனுக்குப் பின்னால் ஓடினோம் _END	5	6
198	Type2	It's been a long time since I've heard anyone use that word.	START_ ஒருவர் அந்த வார்த்தையைப் பயன் படுத்துவதைக் கேட்டு ரொம்ப நாளாகிறது _END	12	10
104	Type2	He can read and write.	START_ அவனுக்கு எழுதப் படிக்கத் தெரியும் _END	5	6
123	Type2	It seems she hates you.	START_ அவள் உன்னை வெறுக்கிற மாதிரி தெரிகிறது _END	5	7

Show 25 per page

Figure 3. After query expansion of English-Tamil

1 to 10 of 50 entries Filter

index	source	malayalam_sentence	tamil_sentence	length_mal_sentence	length_tam_sentence
145	T2	മുൻ വൈരുദ്ധ്യം	START_ குழப்பங்கள் _END	2	3
12	T1	ഈ വേദഗ്രന്ഥത്തിന്റെ അവതരണം പ്രതാപിയും യുക്തിമാനുമായ അല്ലാഹുവിങ്കൽ നിന്നാകുന്നു	START_ இவ்வேதம் யாவரையும் மிகைத்தோனும் ஞானம் மிக்கோனுமாகிய அல்லாஹ்விடமிருந்தே இறக்கியருளப்பட்டது _END	7	9
160	T2	എല്ലായിടത്തും b തെരഞ്ഞെടുക്കുക	START_ எல்லா இடத்திலும் bவைத் தேர்ந்தெடு _END	3	6
26	T1	തൊണ്ടയിൽ അടഞ്ഞു നിൽക്കുന്ന ഭക്ഷണവും വേദനയേറിയ ശിക്ഷയുമുണ്ട്	START_ தொண்டையில் விக்கிக் கொள்ளும் உணவும் நோவினை செய்யும் வேதனையும் இருக்கின்றன _END	6	10
33	T1	എനിക്കറിയില്ല	START_ எனக்கு தெரியாது _END	1	4
65	T2	നിനക്കു മന്ത്രവാദം അറിയാമോ	START_ உமது ஆசான் அதை விளக்கினாரா _END	3	6
111	T2	പ്രധാന ഉപകരണപ്പട്ട	START_ கருவிப்பட்டியைத் தேடு _END	2	4
61	T2	ഇന്നു പ്രായവും ൽ	START_ இந்த நாள் மற்றும் வயது _END	3	6
158	T2	വരി നമ്പർ	START_ வரி எண்களை காட்டு _END	2	5
90	T2	ഇതു ഒരു ശബ്ദ ഫയൽ അല്ല	START_ இது ஒலி கோப்பு இல்லை _END	5	6

Show 10 per page 1 2 3 4 5

Figure 4. After query expansion of Malayalam-Tamil

3.3. Query Translation

A pre-trained Long Short-Term Memory (LSTM) model is used to map the query in the source language into an equivalent query in the language of the target document collection [40]. LSTM based encoder-decoder architecture model is used in this study, which is used to remember long-term and short-term dependencies. A basic form of LSTM consists of two components; an encoder which computes a

representation of source sentences and a decoder which generates corresponding target word in the target languages.

ELRIC bi-lingual parallel corpora are used in the work as the dataset. The two datasets are in the form of CSV file with utf-8 encoding format of both English-Tamil parallel corpora and Malayalam-Tamil parallel corpora. The first dataset consists of 148 English words, 176 Tamil words and total number of equivalent sentences is 201 of both English and Tamil languages. The second data set consists of 127 Malayalam words, 141 Tamil words and total number of equivalent sentences is 172. The model translates English to Tamil and Malayalam to Tamil.

Here pickle format of data is used, which is a python dependency and can serialize our dataset. In this model TensorFlow's built-in data_utils class is used to prepare the data. data_utils class is used to read the data from the directory, pre-process, and format words from both languages. Data is pre-processed with Conversion of all characters into lowercase, removing quotes, remove the all-special characters, remove the all the digits from the documents, then add starting and end markers for all the sentences finally to build the source and target language vocabulary. LSTM based auto encoder-decoder model is used for translation of source to target language. The data set is divided into two such as training and testing, the model split the data into an 80-20 ratio of training and testing. The model optimization parameters include batch size as 40, zero padding as 1, LSTM activation function as softmax, categorical_crossentropy is used as loss function and model optimizer as 'rmsprop' LSTM encoder-decoder model is divided into different layers such as input layer, 2 embedding layers, 2 LSTM layer and output layer.

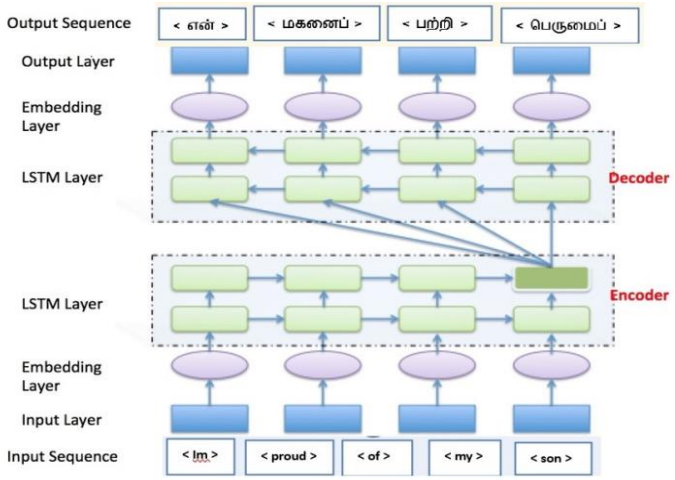

Figure 5. LSTM based auto encoder-decoder architecture

The first layer accepts the input sequences after that input sequence token is will embedded with some learning parameters. Layer2 used for encoding input sequences and Layer3 used for decoding the output sequences. Total number of Parameters is 1598877 and the model is trained for 100 epochs. The layer architectural model summary is given below;

```
Layer (type)                 Output Shape              Param #   Connected to
==================================================================================================
input_1 (InputLayer)         [(None, None)]            0         []

input_2 (InputLayer)         [(None, None)]            0         []

embedding (Embedding)        (None, None, 300)         50100     ['input_1[0][0]']

embedding_1 (Embedding)      (None, None, 300)         53100     ['input_2[0][0]']

lstm (LSTM)                  [(None, 300),             721200    ['embedding[0][0]']
                              (None, 300),
                              (None, 300)]

lstm_1 (LSTM)                [(None, None, 300),       721200    ['embedding_1[0][0]',
                              (None, 300),                        'lstm[0][1]',
                              (None, 300)]                        'lstm[0][2]']

dense (Dense)                (None, None, 177)         53277     ['lstm_1[0][0]']

==================================================================================================
Total params: 1598877 (6.10 MB)
Trainable params: 1598877 (6.10 MB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 6. Summary of LSTM model for language translation

After model training the system translate from source to target such as English to Tamil and Malayalam to Tamil respectively. The evaluation purpose we were randomly selected some sentence from these corpora.

4. RESULTS AND DISCUSSION

The model translated each sentence by sentence with translation equivalence. After translation, accuracy was evaluated using accuracy metric like BLEU. BLEU stands for Bilingual Evaluation Understudy is an n-gram based evaluation metric [41]. BLEU score is used to compare candidate translation of the text with available reference translations. The task of a BLEU score is to compare n-grams of the candidate with the n-gram of the reference translation and count the number of matches. These matches are position-independent. It is expressed as the following equation (1)

$$\text{BLEU} = \text{BP}exp(\sum_{n=0}^{N} w_n log P_n) \quad (1)$$

Where P_n is an n-gram precision that uses n-grams up to length N, w_n is positive weights that sum to one and BP is Brevity Penalty which is computed as following equation (2)

$$BP = \begin{cases} 1, & if\ c > r \\ e^{(1-r/c)}, & if\ c \leq r \end{cases} \quad (2)$$

Where, c is the length of the candidate translation and r is referenced corpus length.

The table consists of sample source sentences and its translation equivalence of LSTM model and Google translate with corresponding BLEU score.

Table 3. English to Tamil LSTM and Google Translation with corresponding BLEU Score

Sentences (SE)	BLEU Score	
	LSTM (LS)	Google Translate (GT)
SE: Where do you keep your passport		
LS: நீ பாஸ்போர்ட்டை எங்கே வைத்திருக்கிறாய்	**1.00**	--
GT: உங்கள் பாஸ்போர்ட்டை எங்கேவைத்திருக்கிறீர்கள்	--	0.79
SE: He can read and write		
LS: அவனுக்கு எழுதப்படிக்கத் தெரியும்	**1.00**	--
GT: அவருக்கு எழுதவும் படிக்கவும் தெரியும்	--	0.84
SE: I'm proud of my son		
LS: என் மகனைப்பற்றி பெருமைப்படுகிறேன்	0.84	--
GT: என் மகனைப்பற்றி நான் பெருமைப்படுகிறேன்	--	**0.93**
SE: What did he say		
LS: அவன் என்ன ஆரம்பித்தான்	0.84	--
GT: அவர் என்ன சொன்னார்	--	**0.88**
SE: Tom has been crying all afternoon		
LS: அவள் மதியம் முழுவதும் அழுது கொண்டேயிருக்கிறான்	0.88	--
GT: டாம் மதியம் முழுவதும் அழுது கொண்டிருந்தான்	--	**0.93**

SE- Sentence, GT- Google Translate, LS- LSTM Translation

Table 4. Malayalam to Tamil LSTM and Google Translation with corresponding BLEU Score

Sentences (SE)	BLEU Score	
	LSTM (LS)	Google Translate (GT)
SE: ആദ്യഡെൽട്ടയി ലേക്ക്പോകുക (Go to the first delta)		
LS: முதல் டெல்டாவிற்குச் செல்க	**1.00**	--
GT: முதல் டெலட்டிற்கு செல்லவும்	--	0.66
SE: അടു ത്തലേഖനം (Next Article)		
LS: அடுத்த கட்டுரை	**1.00**	--
GT: அடுத்த கட்டுரை	--	**1.00**
SE: a യിൽനിന്നു വരികൾതെ രഞ്ഞെടുക്കുക (Select lines from a)		
LS: a விலிருந்து வரிகளைகளை தேர்ந்தெடு	**0.94**	--
GT: a இலிருந்து வரிசைகளைத் தேர்ந்தெடுக்கவும்	--	0.84
SE: ഒഴിഞ്ഞ ഒപ്പ് (Empty signature)		
LS: வெற்று கையொப்பம்	**0.99**	--
GT: காலி ஒப்	--	0.63
SE: മുൻവൈരുദ്ധ്യത്തി ലേക്ക്പോകുക (Go to previous conflict)		
LS: முந்தைய டெல்டாவிற்குச் செல்க	**0.84**	--
GT: முந்தைய மோதலுக்குச் செல்லவும்		0.63

SE- Sentence, GT- Google Translate, LS- LSTM Translation

Table 3 and 4 show the sentences translated from English to Tamil and Malayalam to Tamil with two the translation models and their corresponding BLEU scores. These results are obtained on the same document collections as the above the given corpus. The five sentences are selected from each corpus. The range of BLEU score is defined as from 0.00 to 1.00, where 0.00 represent is translation accuracy is less and 1.00 represent is accuracy is high. The score values only depend on matching number of words in adjacent position. The average BLEU score is estimated the following two tables 4 and 5 are given below;

Table 5. Average BLEU Score of English to Tamil with sample sentences

Test sentences from English to Tamil data set at random selection	BLEU Score	
	LSTM	Google Translate
English-to-Tamil 1	**1.00**	0.79
English-to-Tamil 2	**1.00**	0.84
English-to-Tamil 3	0.84	**0.93**
English-to-Tamil 4	0.84	**0.88**
English-to-Tamil 5	0.88	**0.93**
Average BLEU Score	**0.912**	**0.874**

Table 6. Average BLEU Score of Malayalam to Tamil with sample sentences

Test sentences from Malayalam to Tamil data set at random selection	BLEU Score	
	LSTM	Google Translate
Malayalam-to-Tamil 1	**1.00**	0.66
Malayalam-to-Tamil 2	**1.00**	**1.00**
Malayalam-to-Tamil 3	**0.94**	0.84
Malayalam-to-Tamil 4	**0.99**	0.63
Malayalam-to-Tamil 5	**0.84**	0.63
Average BLEU Score	**0.954**	**0.752**

Tables 5 and 6 shows that average BLEU of both LSTM based encoder-decoder and Google translation. An evaluation was conducted with selected five random sentences form these two datasets. The selected sentences translated with two translation model such as LSTM and Google translate. Each sentence compared with some set of possible reference translated sentences prepared by language experts. Finally comparing values measured using n-grams with size 2 (i.e. bi-gram). Average BLEU score of sentences translated from English to Tamil in LSTM is 0.912 and Google translate is 0.874. The LSTM got the highest score among these two models. Similarly, the average score of Malayalam to Tamil selected sentences using LSTM is 0.954 and Google translate is 0.752. Each translated sentence is compared with 5 to 10 references. Finally, the paper compared performance level of related models with BLEU score accuracy.

This section provides a summary and comparative analysis of various approaches proposed by different authors in the field. Table 7 illustrates that there has been limited research conducted specifically on Tamil and Malayalam languages, and majority of works focusing on English in connection with other regional languages. Many authors have collected bilingual corpora for translation and training purposes, emphasizing the significance of translation in every CLIR model.

BLEU score metric is based on estimating the number of n-grams matched between the target and references sentences. While document translation can be performed both online and offline, query translation relies solely on online methods. Previous studies have predominantly concentrated on both query and document translation. When comparing our proposed model with others, BLEU Score is considered as evaluation parameter to in this study. Our model exhibits the highest accuracy rate at 0.95 when compared with other translation models. Each model defines specific datasets with varying data sizes, methodologies, and results of CLIR models.

Table 7. Performance of several techniques proposed in CLIR models

Authors and Year	Languages	Data size	Tasks	Models	BLEU Score
Ibrahim Kassa et al. (2021)	Amharic and Arabic	Ethiopian Language Researchcentre (ELRC) annotated corpus-2,10,000 words from news domain	Amharic-Arabic cross language information retrieval using language modelling	Pre-trained Neural Machine Translation with Parts-of-Speech Tagging (POS)	Amharic- 0.88 Arabic-0.94
Zeeshan et al. (2020)	Chinese and Urdu	66,000 Chinese-Urdu parallel corpus	Urdu Word base dictionary translation using Neural Machine Translation	Deep learning model for machine translation	LSTM- 0.94 Transformer-0.77 Google translate- 0.74
Himanshu et al. (2020)	English-Tamil and English-Malayalam	EnTam V2.0, UMC005 and Opus. News and Cinema Domain	Sequence to sequence NMT machine translation system for low resourced languages	NMT techniques using Word-embedding along with pre-trained Byte-Pair Encoding (BPE)	English-Tamil: 0.94 English-Malayalam: 0.90
Mary Priya et al. (2010)	English and Malayalam	Online Malayalam newspaper and magazines	Translation between English and Malayalam using Statistical approach	Statistical Machine Translation	English to Malayalam-0.69
Sakthi Vel et al. (2023) (Proposed Work)	**Tamil and Malayalam**	**European Research Infrastructure Consortium (CLARIN-ERIC) Corpus- 373 Sentences**	**Review the collected CLIR works and proposed a model for translation of Tamil and Malayalam**	**Translation model using LSTM (Proposed Work)**	**LSTM- 0.95** **Google Translate-0.752**

5. CONCLUSION

The proposed LSTM-based translation model using two bilingual corpora outperforms Google Translate with BLEU score evaluation metric. The LSTM based auto encoder-decoder was trained with selected random sentences from the given corpora. The Average BLEU score of English to Tamil is 0.912 and Malayalam to Tamil as 0.954 whereas Google translate gained 0.874 and 0.752 respectively for the two datasets. Google translation model is based on conditional probability, whereas the LSTM model based on sequence-to-sequence relationships mapping with adjacent words.

The results obtained are compared with state of art methods and results reveal that the LSTM model gain and outperformed with highest accuracy. The overall performance of each model depends on different techniques, language complexity, data size and similar factors.Future enhancements may involve increasing dataset volume, integrating grammatical rules, incorporating parts of speech tagging, and implementing advanced document retrieval techniques to refine the CLIR framework for more effective crosslanguage information retrieval systems.In the future, the possibility to improve the translation results for different domains of these languages can also be explored.

REFERENCES

[1] Verma, Nitin., Arora, Suket., Verma, Preeti.: Cross-Language Information Retrieval on Indian Language: A Review. IITM Journal of Management and IT, Vol. 8, Issue 1, pp. 63-66, (2017).
[2] Dolores, Maria., and Lobo. Olvera, "Cross-Language Information Retrieval on the Web", IGI Global, pp.704-719, 2009.
[3] S. Pourmahmoud and M. Shamsfard, "Semantic Cross-lingual Information Retrieval," 2008 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, 2008, pp. 1-4, doi: 10.1109/ISCIS.2008.4717868.
[4] Litschko. Robert, Glavas. Goran, Ponzetto. Simone Paolo, and Vulic. Ivan, "Unsupervised Cross-Lingual Information Retrieval using Monolingual Data Only:, ACM, pp.1-5, arXiv: 1805.00879v1 [cs.CL.], 2018.
[5] Zhuhadar, Leyla., Nasraoui, Olfa.: Evaluating a Cross-Language Semantically Enriched Search Engine. Proceedings of 2010 Seventh International Conference on Information Technology, pp. 1074-1079, 978-0-7695-3984-3/10, (2010).
[6] Zhuhadar, Leyla., Nasraoui, Olfa., Wyatt, Robert., Romero, Elizabeth.: Multi-Language Ontology-based Search Engine. Proceedings of 2010 Third International Conference on Advances in Computer-Human Interactions, pp. 13-18, 978-0-7695-3957-1/10, (2010).
[7] N. Jian-Yun, "Cross-Language Information Retrieval", IEEE Computational Intelligence Bulletin, Vol.2, No.1, pp. 19-24, 2003.
[8] M. N. Asim, M. Wasim, M. U. Ghani Khan, N. Mahmood and W. Mahmood, "The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval," in IEEE Access, vol. 7, pp. 21662-21686, 2019, doi: 10.1109/ACCESS.2019.2897849.

[9] A. Mustafa, T. John, and O. Michael, “Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base”, Polibits (40), pp.13-16, 2009.
[10] S. Pourmahmoud and M. Shamsfard, "Semantic Cross-lingual Information Retrieval," 2008 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, 2008, pp. 1-4, doi: 10.1109/ISCIS.2008.4717868.
[11] Sharma. Monika, and Morwal, Sudha, “Refinement of search results using cross lingual reference technique”, International Journal of Advanced Research in Computer and communication Engineering, Vol. 3, Issue 12, pp. 8692-8695, ISSN: 2278-1021, 2014.
[12] Gupta. Parul, and Sharma. AK, “Context based Indexing in Search Engines using Ontology”, International Journal of Computer Applications, Vol. I, No.14, pp.49-52, ISSN: 0975-8887, 2010.
[13] B. A. Kumar, "Profound Survey on Cross Language Information Retrieval Methods (CLIR)," 2012 Second International Conference on Advanced Computing & Communication Technologies, Rohtak, India, 2012, pp. 64-68, doi: 10.1109/ACCT.2012.91.
[14] Zeeshan, Jawad and M. Zakira, ”Research on Chinese-Urdu Machine Translation Based on Deep Learning,” Journal of Autonomous Intelligence, 2020, Vol. 3, Issue 2, pp. 34-44. Doi:10.32629/jai.v3i2.279.
[15] K. Aditi, K. Hemant, P.Shashi, K. Ajai and D. Hemant, “Evaluation and Ranking of Machine Translated Output in Hindi Language using Precision and Recall Oriented Metrics,” International Journal of Advanced Computer Research, 2014-March, Vol. 4, Issue. 14, pp. 54-59.
[16] P. L. Nikesh, S. M. Idicula and S. David Peter, "English-Malayalam Cross-Lingual Information Retrieval- an experience," 2008 IEEE International Conference on Electro/Information Technology, Ames, IA, USA, 2008, pp. 271-275, doi: 10.1109/EIT.2008.4554312.
[17] Kassa, Ibrahim Gashaw., Shashirekha, H.L.: A^2 CLIR: Amharic-Arabic Cross Language Information Retrieval Using Language Modeling. UGC Shodhganga Inflibnet, https://shodhganga.inflibnet.ac.in/handle/10603/380186, (2021).
[18] Thenmozhi, D., Aravindan, Chandrabose.: Ontology-based Tamil-English Cross-lingual Information Retrieval system. Indian Academy of Sciences, Sadhana, 43:157, pp. 3-14, https://doi.org/10.1007/s12046-018-0942-7, (2018).
[19] PV. Vidya, PC. Raj, V. Reghu, and Jayan, “Web Page Ranking Using Multilingual Information Search Algorithm: A Novel Approach”, ICETEST-2015, Procedia Technology-Elsevier Publication, pp.1240-1247, doi: 10.1016/j.protcy.2016.05.102, 2015.
[20] Shree. KV, Saviya. E, Umamaheswari, J. Balaji., Geetha, TV., and Parthasarathi, Ranjani, “Conceptual Based Search Engine (CBSE) system for Tamil and English”,TaCoLa Lab, CEG, Anna University, Chennai. pp.105-111.
[21] E. Katta and A. Arora, "An improved approach to English-Hindi based Cross Language Information Retrieval system," 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, India, 2015, pp. 354-359, doi: 10.1109/IC3.2015.7346706.
[22] Mayanale. Savita C, and Pawar. SS, “Marathi-English CLIR using detailed user query and unsupervised corpus-based WSD”, Int. Journal of Engineering Research and Applications, Vol.5, Issue 6, pp.86-91, ISSN: 2248-962, 2015.
[23] S. Saraswathi, A. Siddhiqaa, K. Kalaimagal, and M. Kalaiyarasi, “Bi-Lingual Information Retrieval System for English and Tamil”, Journal of Computing, Vol.2, Issue 4, pp. 85-89, ISSN 2151-9617, 2010.
[24] Reddy, Mallamma V., Hanumanthappa, M., and Kumar, Manish, “Cross Lingual Information Retrieval Using Search Engine and Data Mining”, ACEEE Int. J. on Information Technology, Vol.01, No.02, pp.10-13, 2011.
[25] Chandra, Ganesh., Dwivedi, Sanjay Kumar.: Applying Query Expansion in Cross Lingual IR (Hindi-English) for Relevancy Improvements. UGC Shodhganga- Inflibnet, https://shodhganga.inflibnet.ac.in/handle/10603/260610. 2017.
[26] A. Mustafa, T. John, and O. Michael, “Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base”, Polibits (40), pp.13-16, 2009.
[27] P. Bajpai, Pratibha, V. and, Parul, “Cross Language Information Retrieval: In Indian Language Perspective”, International Journal of Research in Engineering and Technology. Vol. 03, Special Issue. 10, pp.46-52, 2014.
[28] Sharma. Monika, and Morwal, Sudha, “Refinement of search results using cross lingual reference technique”, International Journal of Advanced Research in Computer and communication Engineering, Vol. 3, Issue 12, pp. 8692-8695, ISSN: 2278-1021, 2014.
[29] Litschko. Robert, Glavas. Goran, Ponzetto. Simone Paolo, and Vulic. Ivan, “Unsupervised Cross-Lingual Information Retrieval using Monolingual Data Only:, ACM, pp.1-5, arXiv: 1805.00879v1 [cs.CL.], 2018.
[30] Song, Xiang., Zhou, Jialiang., Kimura, Fuminori., Maeda, Akira.: A Japanese-Chinese Cross-Language Entity Linking Method with Entity Disambiguation Based on Document Similarity. International Journal of Knowledge Engineering, Vol.2, No.3, pp.122-127 (2016).
[31] Zhuhadar, Leyla., Nasraoui, Olfa.: Evaluating a Cross-Language Semantically Enriched Search Engine. Proceedings of 2010 Seventh International Conference on Information Technology, pp. 1074-1079, 978-0-7695-3984-3/10, (2010).
[32] S. J. Shim, "Using Cross-Language Information Retrieval Methods for Bilingual Search of the Web," International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), Vienna, Austria, 2005, pp. 19-24, doi: 10.1109/CIMCA.2005.1631439.

[33] O. Attia, M. Azmy, Emeira. Ahmed Abu., Azzouni, Karim El., Hussein, Omar., El-Makky, Nagwa M., N. and, Khaled, “Using Deep Learning in Arabic-English Cross Language Information Retrieval”, Egyptian Information Technology Industry Development Agency (ITIDA), pp.1-8, 2012.
[34] M. Nassirudin and A. Purwarianti, "Indonesian-Japanese term extraction from bilingual corpora using machine learning," 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2015, pp. 111-116, doi: 10.1109/ICACSIS.2015.7415180.
[35] Zhuhadar, Leyla., Nasraoui, Olfa., Wyatt, Robert., Romero, Elizabeth.: Multi-Language Ontology-based Search Engine. Proceedings of 2010 Third International Conference on Advances in Computer-Human Interactions, pp. 13-18, 978-0-7695-3957-1/10, (2010).
[36] Azarbonyad, H., Shakery, A., Faili, H. (2013). Exploiting Multiple Translation Resources for English-Persian Cross Language Information Retrieval. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) Information Access Evaluation. Multilinguality, Multimodality, and Visualization. CLEF 2013. Lecture Notes in Computer Science, vol. 8138. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40802-1_11.
[37] https://www.clarin.eu/resource-families/parallel-corpora
[38] https://github.com/Kartikaggarwal98/Indian_ParallelCorpus
[39] Vel, Sakthi. And R, Priya. (2022). "Text Pre-Processing Methods on Cross Language Information Retrieval," 2022 International Conference on Connected Systems & Intelligence (CSI), Trivandrum, India, pp. 1-5, doi: 10.1109/CSI54720.2022.9923952.
[40] https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/
[41] https://www.geeksforgeeks.org/nlp-bleu-score-for-evaluating-neural-machine-translation-python/
[42] Choudhary, Himanshu., Rao, Shivansh., and Rohilla, Rajesh. “Neural Machine Translation for Low-Resourced Indian Languages”, proceedings of the 12th conference on Language Resources and Evaluation, pp.3610-3615. 2020.
[43] Sebastian, Mary Priya., Kurian L, Sheena., and Kumar, G.Santhosh. “English to Malayalam Translation: A Statistical Approach”, A2CWiC '10: Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India. September 2010. Article No.: 64. Pages 1–5. https://doi.org/10.1145/1858378.1858442.

BIOGRAPHY OF AUTHORS

Dr. Priya R. is currently working as Assistant Professor, Department of Computer Science, Government College Kariavattom, Kerala. She completed her PhD from University of Kerala. She has more than 20 years teaching experience. She has research publications in several International Journals and conferences. Her research interest are Modeling of constrained complex systems, Soft computing, Machine Learning, Object Oriented Programming and Language Computing.

Sakthi Vel S., completed his Master degree from University of Kerala. Currently doing PhD from Centre for Development of Imaging Technology (C-DIT), University of Kerala. He has more than seven years of teaching experience at university level. His research interest are Natural Language Processing and Language Computing.