Computing and Informatics, Vol. 40, 2021, 628–647, doi: 10.31577/cai_2021_3_628

LEARNING TO TRANSLATE KANNADA AND ENGLISH QUERIES FOR MIXED SCRIPT INFORMATION RETRIEVAL

B. S. Sowmya Lakshmi

Department of Machine Learning
B. M. S. College of Engineering
Bangalore, Karnataka
e-mail: sowmyalakshmibs.mel@bmsce.ac.in

B. R. Shambhavi

Department of Information Science and Engineering
B. M. S. College of Engineering
Bangalore, Karnataka
e-mail: shambhavibr.ise@bmsce.ac.in

Abstract. Due to increase in the availability of numerous languages in the Web, cross language information retrieval is one of the happening issues in the field of natural language processing and information retrieval. Nowadays, people are habituated to combine two or more language words during oral or written discourse. Speakers have also employed intermixing of different languages and scripts in digital media while querying, blogging and on social media platforms. The way of representing two different language words of an utterance in their native scripts is known as mixed scripting. In the present work, we attempted to translate mixed script queries of Kannada and English languages into monolingual queries. We proposed three approaches for translation by constructing bilingual dictionary, word embeddings and Google translate. The proposed method outperforms the conventional dictionary based approach, when word embeddings were combined with the translations learnt from Google Translate and Dictionary.

Keywords: Code mixing, mixed script queries, cross language information retrieval, machine translation

1 INTRODUCTION

1.1 Information Retrieval (IR)

The term "Information Retrieval" was first devised by Calvin Mooers in 1950s. Later on many researchers focused on IR in the mid of 1990s.

According to Manning et al., "Information Retrieval" refers to the technology of "finding material (usually documents) of an unstructured nature (usually text) that satisfies information need from within large collections (usually stored on computers)". The term "material" can be understood in many folds as tweets, videos, music, books, images, documents etc. In this study, we restrict ourselves to text data. Basic terminology behind IR are:

Corpus: A large repository of documents stored on single or multiple computers.

Information need: A topic about which user wants the information, often referred as query.

Relevance: Few of the documents in the corpus might contain topics relevant to information need.

There exist three flavours of IR based of the degree of retrieval as follows.

Web search: WWW is a huge repository of contents which can be searched with the aid of search engines like Google, Bing etc.

Enterprise search: It can also be called Intranet search where search for documents is confined inside a particular organization or company.

Personal search: This search is restricted to one's personal computer where the user search required file stored in his computer. The collection is typically a set of files on a personal computer of the user.

1.2 Cross Language Information Retrieval (CLIR)

As more digital information is made available, the Web continues to be the foremost channel for communication and the largest data repository. Besides the large number of English speaking users, dominance of English on the Web is caused also by the fact that several organizations create English versions of their websites (besides those in their native languages) and of their broad business needs, probably to be widely accessible. Governments around the world also imposed English as a formal language, to some extent, in their educational and governmental spheres. As a result, English was, and still is, the most dominant language for scientific articles, lexicons, dissemination of information and different types of knowledge. However, there exist growth of non-English languages on the Web, as some governments enforce that national corporations and organizations publish some material like people's heritage, geographical data and educational technical material in native languages.

Accordingly, more and more pages on the Web are written in different languages. This resulted in globalized information and a large number of resources that are very much diverse and in a multitude of languages. This feature makes the Web essentially cross-lingual and/or multilingual. But, this linguistic multiplicity and moving towards an international community should no longer be a barrier for accessing information, regardless of its language, on the Web. When users need to search in any language for a particular topic, the search results should no longer be restricted to the native languages of those users. For such users, CLIR provides a solution. In CLIR, users are able to obtain relevant information (document sets) in a language that is different from the language they used in their information need requests (queries). For example, a user may type his/her query in Kannada, a South Indian language, but relevant document sets retrieved are in English or any other language. CLIR system is more complex than traditional monolingual IR system as CLIR also includes a Translation phase.

In query translation approach, query in the source language is translated into the language in which documents are to be retrieved (target language). Machine Translation (MT) is the task of automatically converting the sentences in one natural language into another, preserving the meaning of the input text and producing fluent text in output language. The main objective is to fill up the language gap between two different languages speaking people, communities or countries. The goals of proposed MT system are as follows:

1. In the proposed approach, as input is a Mixed Script query adoption of POS tagging would make the translation process fruitful. Unlike, conventional MT systems, we followed word by word translations by ignoring syntax structure of the respective language. So, if POS of each word in the Mixed Script query is known then translation could be performed in accordance with the POS tags.
2. To develop a bilingual dictionary of Kannada and English Languages.
3. To develop a MT system to translate Kannada – English mixed script queries into monolingual Kannada and English queries.

Handling Indian languages is a challenging task as they differ in morphology and semantic features from English. Even though, the worldwide web is a host to numerous languages, statistics shows that English holds the major share of documents and usage. Which results in creation of Mixed Script space, having documents and queries in single or multiple languages in one or more scripts. IR in Mixed Script space can be called Mixed Script IR (MSIR). MSIR is more challenging than IR as it involves understanding and matching of queries written in two or more scripts with the documents in either of the scripts.

There have been several studies on CLIR including Indian languages. In a CLIR setup, language of the query and retrieved documents are different. MSIR deals with querying in more than one language to retrieve documents in one or more languages. In either case, the documents and the query are written in their native scripts. This

article intends to familiarize the issue of MSIR for Kannada-English mixed query terms. Present state of the art systems are unable to process Mixed Script Queries due to the lack of resources such as transliterated dictionaries and MT systems. Semantic search for Mixed-Script query is still an unsolved problem and it increases in many folds when applied on web search. Adequate tools are not available to process queries having Mixed-Script terms.

The major contributions of this paper are:

- To present the concept, formal definition of MSIR from web for Indian languages, particularly Kannada-English bilingual texts.
- To create a POS tagger for Kannada words.
- To demonstrate how difficult the MSIR problem is and where existing IR techniques fail when applied on such data.

The remaining sections of this paper are arranged as follows. Section 2 describes prior research in this area. Proposed method for translating Mixed Script query is described in Section 3. The results obtained are briefed in Section 4. Section 5 communicates conclusion of the proposed method.

2 LITERATURE SURVEY

Though MSIR has achieved very modest consideration, many laterally correlated tasks like CLIR and transliteration reveals few problems of MSIR. Whereas languages like Chinese and Japanese follow more than one script [1], they might not come across the actual difficulty of the MSIR as they abide by benchmark rules for script writing and spelling. However, this is not true in the case of Indian languages. For instance, in Romanization of Kannada words, there exists no such rules resulting in great number of discrepancies. Furthermore, these Romanized words are combined with English words making difficult to identify transliterated text.

In CLIR queries are translated to the language of the document set. However, out of vocabulary words like Named Entities need to be transliterated rather than translated. There exist no standard rules for mapping alphabets of Indian languages to English or vice versa. This has led to lot of discrepancies in developing a transliteration model [2]. Most of the researchers have highlighted the difficulties in developing transliterated language models for Indian languages in web search [3]. Researchers emphasize on this issue in Hindi Song Search system in Latin script [4, 5]. They focused handling transliterated word pairs matching while crawling song lyrics from various websites from the web. Edit-distance is one among most familiar methods for matching word pairs. Authors in [6] and [7] have followed this method for English-Telugu and Tamil-English language pairs, respectively. Authors in [8] proposed a method to normalize transliterated text using combination-based approach in which a statistical stemmer is used to delete commonly used suffixes along with rules to map spelling variants. An equivalent system that handles both stemming and conversion of grapheme to phoneme was used in [9] to build a standalone search

engine for ten Indian languages. Even though, there are few substantial works present in the field of handling variants and normalization of transliterated text, in practice the process of MSIR is largely ignored.

Gupta et al. [10], analysed query log data of most familiar Bing search engine, to evaluate the significance of their MSIR system. They projected a deep learning paradigm to match mixed- script terms and handle variations in spelling. Method significantly achieved better results when judged with Naive Bayes model by 12 % and 29 % increase in Mean Recall and Mean average precision value.

Pathak et al. [11] attempted to create Automatic Parallel Corpus Creation for Hindi English News Translation Task. Authors developed parallel corpus from comparable corpus crawled from the web from various sources. Quality of the parallel corpus created was analysed by Gestalt Pattern Matching, Hamming Distance and Levenshtein Distance algorithm to calculate sentence matching between Hindi – English sentences. Li et al. [12] developed a Neural Machine Translation (NMT) system, which learns a general network as usual, and then fine-tunes the network for each test sentence. The fine-tune work was done on a small set of the bilingual training data that was obtained through similarity search according to the test sentence. Similarity among sentences were calculated using Levenshtein distance, average word embedding and hidden states of the encoder in NMT measures. Authors observed that performance of Levenshtein distance based similarity was better than other two measures.

Another well known metric used for evaluating machine translation is Evaluation of Translation with Explicit Ordering (METEOR) [13]. Dunđer et al. [14] proposed a machine translation for poetry and a low resource language pair, such as Croatian-German. The authors collected data set that contained the works of a contemporary poet of the Croatian language and the translations of his poems in German. Results were evaluated through BLEU, METEOR, RIBES and Character metrics. An English to Urdu and Hindi translation system was developed using Neural network and translation rules by Khan and Usman [15]. System was evaluated using n-gram, BLEU, METEOR, presion and F-measure scores. METEOR score achieved was 0.7956 for Urdu and 0.8083 for Hindi.

The necessity to recognize and process Indian language scripts is in demand as nearly 50 % of the Indian population use internet daily (according to statistics). Indian Language Technology Proliferation and Deployment Centre (TDIL-DC) has provided phonetic keyboard input is support for all Indian languages. However, POS tagging on Indian languages and especially on Dravidian Languages is quite a difficult task due to the unavailability of annotated data for these languages. Various techniques have been applied for POS tagging in Indian languages. Gadde and Yeleti [16] used morphological features with Hidden Markov Model (HMM) tagger and obtained 92.36 % for Hindi and 91.23 % for Telugu. The Hindi POS tagger used Hindi Treebank 3 of size 450 K. Ekbal and Bandyopadhyay [17] used Support Vector Machine (SVM) for POS tagging in Bengali obtaining 86 % accuracy. The POS tagging in morphologically richer Dravidian Indian languages has always posed a great challenge for researchers. Malayalam is a highly agglutinative language

in the Dravidian family. Sandhi splitting or word segmentation between conjoined words should precede the POS tagging to find word boundaries.

Devadath and Sharma [18] explored the significance of Sandhi splitting on shallow parser and built a POS tagger using Conditional Random Field (CRF). Their POS tagger performed well with 90.45 % accuracy. Antony et al. [19] used SVM with lexicon to obtain 94 % accuracy. A semi-supervised pattern-based bootstrapping technique was implemented by Ganesh et al. [20] to build a Tamil POS Tagger. Their system scored 87.74 % accuracy on 20 000 documents containing 271 K unique words.

Due to the scarcity of quality annotated data very little work has been done on Kannada language. Kannada language has a free form of word arrangement in a sentence which makes POS tagging task for Kannada rigid. Most of the recent works in POS tagging on Kannada have been experimented only with traditional ML techniques like HMM, CRF or SVM. One of such noticeable works was proposed by Shambhavi and Kumar [21]. Authors focused on assignment of POS tags for every word belonging to input Kannada language sentences using machine learning algorithms like second-order HMM and CRF. They have used EMILLE corpus which has 51 269 words as train data, and 2 932 words as test data. Authors were able to achieve accuracies of 79.9 % and 84.58 % for HMM and CRF methods, respectively.

Graves and Schmidhuber [22] proposed a POS tagger for Kannada language by applying CRF with corpus consisting 80 000 words. They followed TDIL tags for training and testing the system. They obtained an accuracy of 92.4 % for POS tagging.

3 CONTRIBUTION

3.1 Corpora Extraction

MT is one of the well known NLP applications. In the recent years, MT systems are built based on Neural network approach [23], parallel data or with the aid of bilingual dictionaries. It is hard to find a machine readable dictionary for resource scarce language like Kannada. Bilingual dictionaries are usually built using sentence aligned parallel text corpus. But, latest advances in developing a bilingual dictionary is using comparable corpora [24]. Wikipedia is a well- known comparable corpora, we used Wikipedia for the construction of bilingual dictionary of Kannada and English language pair.

Wikipedia contains wide range of articles in different languages and several link statistics amongst articles. It is being utilized as corpora in various NLP tasks fruitfully. Pages on Wikipedia connect to equivalent pages in other languages on similar topic via interlanguage links. For instance, there exist an interlanguage link between English article “Telephone” to the corresponding Kannada article, as depicted in Figure 1.

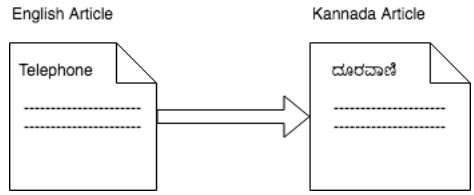

Figure 1. Interlanguage link example

Titles of the majority of the articles which are associated by an interlanguage links are translations of each other. Even though interlanguage links are accurate, there exists some extent of discrepancies as links are generally created manually. It is observed that, in addition to article titles, text inside the articles also share parallel contents to the great extent. Grounded on the above observation, we used interlanguage link to collect Kannada-English comparable corpus from Wikipedia. Steps adapted to develop comparable corpora of 19 263 articles of English and Kannada from Wikipedia are as follows:

Step 1: Kannada and English latest Wikipedia database dump was downloaded from `http://download.wikimedia.org` using a python script.

Step 2: Articles in English which have Kannada interlanguage link were downloaded, followed by the extraction of linked Kannada articles.

Step 3: Paragraphs under each heading are assumed to be related and those which contained general information are retained to ensure comparability.

Step 4: Extracted articles are cleaned by removing unrelated words and superlinks.

3.2 Bilingual Dictionary Creation

Proposed method to create bilingual dictionary, assumes that there exists a correlation amongst the patterns of word-co-occurrence across languages. However, it only requires a medium set of comparable documents which are pre-aligned documents with similar topics.

1. Generating Named Entity Dictionary

   Named Entities (NEs) are the names of persons, organizations, companies etc., i.e., during translation NEs should be transliterated rather than translation. Most of the conventional dictionaries do not have NEs. We took advantage of these NEs to locate comparable sentences in both Kannada and English documents. Also, these NE mapping helped us to find similar sentences across sections. So, to begin with, we tried to map NE in similar articles of both languages. A list of every NE in each English article in the downloaded corpus was created. NE recognition of English words was performed using built

in Stanford NE tagger in Python. Using the combination based transliteration algorithm [25] identified NEs were transliterated to Kannada script. The resulted transliterated NEs in Kannada script were searched and extracted in corresponding Kannada article to match similar sentences. Levenshtein distance algorithm was implemented to perform string matching of transliterated NE and the corresponding NE in Kannada article. Thus, a list of NEs in English articles and its corresponding mapping in Kannada articles was built and appended to our bilingual dictionary. The sentences which contained NE in English and its corresponding Kannada article were short-listed to find to obtain word level association (mappings).

2. Generate Title Dictionary

   Comparable corpus consisted of text related to similar topics but in distinct languages and authored by different authors. Therefore, the article contents may not be precise translations, but they convey information on similar topics. However, titles of such documents are perfect candidates of dictionary entries. To begin with, document title pairs of source and target languages were aligned and preprocessed to remove special characters and numerals. Title pairs were appended to the dictionary, forming a seed dictionary of title pairs. As observed, sub headings of source and target documents may not be same as they are written by different authors. Based on the initial dictionary constructed, related sections of articles in both English and Kannada were found. Sentences which were parallel to some extent are mined from these related sections. Most frequent words in these sentences were appended to the existing dictionary list by calculating word level similarity. Word level similarity was calculated using Pearson correlation coefficient which provided score, where every word in Kannada language gets a score for words in English language. These words were sorted based on their scores to get the best related words in Kannada language for each English word. Algorithm 1 describes generation of title mappings, sub heading mappings and word pair mappings.

   Finally, NE dictionary, title mappings, sub heading mappings and word pair mappings are combined to form a bilingual English–Kannada dictionary.

3. Results

   Dictionary created by proposed method has been evaluated using precision metrics. Precision (P) is the fraction of sum of appropriately (N) translated word pairs to total (T) number of translations in the dictionary which is used to moderate accuracy.

   Bilingual dictionary generated was evaluated manually and their respective precision scores are shown in Table 1.

Algorithm 1 Generate title mappings, sub heading mappings and word pair mappings

```
for all En-document in English-corpus do
    En-title ← Title of En-document
    Ka-document ← corresponding Kannada document in Kannada-corpus
    Ka-title ← Title of Ka_document
    if (En-title, Ka-title) not present in Dict then
        Dict ← Dict U (En-title, Ka-title)
    end if
    for all (En-subheading, Ka-subheading) do
        score-map ← Pearson correlation coefficient(En-subheading,
            Ka-subheading)
        while score-map is not empty do
            (En-subheading, Ka-subheading) ← max-ScoreEntry(score-map)
        if (En-subheading, Ka-subheading) not present in dictionary then
            Dict ← Dict U (En_subheading, Ka_subheading)
            remove all other entries from score-map
        end if
        end while
    end for
    for all partial parallel sentences do
        remove stop words
        Add co-occuring word pairs to score map
        while score-map is not empty do
            (En-word, Ka-word) ← max-ScoreEntry(score-map)
        if (En-word, Ka-word) not present in dictionary then
            Dict ← Dict U (En_word, Ka_word)
            remove all other entries from score_map
        end if
        end while
    end for
end for
```

Phase	Tokens	Precision
Gathering NEs	33 000	0.76
Gathering Title heading	23 398	0.89
Gathering Subheading mapping	1 362	0.86
Co-occuring words	16 785	0.65
Overall	77 545	0.79

Table 1. Precision scores

3.3 Query Translation Process

The wholeness of an IR paradigm stays in its capability to figure out the proper meaning of the input queries before search. In contrast to regular IR, MSIR need a translation system either by human or machine. The proposed Query translation approach ensures to convert user query into document language before retrieval. The proposed approaches to translate mixed script Kannada English queries to monolingual queries are following.

1. Dictionary based translation with POS tagging

   (a) POS Tagging
   In the proposed approach, as input is a Mixed Script query adoption of POS tagging was identified as a fruitful step in the translation process. Unlike, conventional MT systems, we followed word by word translations by ignoring syntax structure of the respective language.
   Input Mixed Script query contains both Kannada and English words in their respective scripts. POS tagging of English words was performed by in built Stanford POS tagger. Kannada words were tagged with BiLSTM-CRF neural network approach, which yielded accuracy of 92 %.

   (b) Translation
   Input query was translated to monolingual Kannada and English queries with the help of bilingual dictionary constructed. Each POS tagged English word in input query was translated to Kannada using dictionary and NEs were transliterated to Kannada script. Thus, forming input query in Kannada language. Query in Kannada language was translated to English using bilingual dictionary forming an English query. All translations were word to word without considering the syntactic structure of the respective languages.

2. Word Embedding (WE) + dictionary

   We found that dictionary-based method fail to translate words which do not have translations. Word Embeddings were adopted to handle such query terms. We trained the word2vec package for both the Kannada and English monolingual documents of comparable corpus obtained from Wikipedia dumps. We used the Continuous Bag of Words (CBOW) model with a window size of 5 and output vector of 300 dimensions with other default parameters set.

   Given an mixed script query as input, each English word in the query translations were taken from the bilingual dictionary, if a translation exists. If not, it is transformed into vector to find similar vector embeddings from corpus, and then translation of a English word of input query to Kannada is performed. Thus, input mixed script query is translated into a monolingual Kannada query. The above technique is followed to translate Kannada words in the input query to English to form monolingual English query.

3. Dictionary +WE+ Google translate

   In this technique a hybrid method is followed by combining dictionary-based method, WE and Google translation. If the translations for input query word does not exist either in the dictionary or in the WEs, then the words were translated using Google translation.

4 RESULTS

We used Anaconda with Python 3 version to build all translation models. We used NLTKs Bilingual Evaluation Understudy (BLEU) Score and Metric for METEOR to evaluate translation performance. BLEU is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another based on n-gram precision. Whereas, METEOR metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. We tested translation paradigm using mixed script queries on current trending topics from Google trends, newspaper headlines and YouTube search queries.

1. Dictionary Based Translation

   (a) English to Kannada translation to form monolingual Kannada Queries
   Table 2 and Figure 2 portrays sample dictionary-based English to Kannada translation and BLEU scores for sample queries, respectively.

Kannada English Input Mixed Script Query	**Translation in Kannada**	**Translations (dictionary)**	**BLEU**	**METEOR score**
ವಿರುಷ್ಕಾ grand reception	ವಿರುಷ್ಕಾ ಅದ್ದೂರಿ ಆರತಕ್ಷತೆ	ವಿರುಷ್ಕಾ grand ಆರತಕ್ಷತೆ	0.81	0.63
where ಕಬ್ಬಿಣದ ಕಂಬ located in india	ಕಬ್ಬಿಣದ ಕಂಬ ಭಾರತದಲ್ಲಿ ಎಲ್ಲಿದೆ	ಎಲ್ಲಿ ಕಬ್ಬಿಣದ ಕಂಬ located in india	0.45	0.45
ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ live telecast	ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ ನೇರ ಪ್ರಸಾರ	ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ live telecast	0.57	0.59

Table 2. Sample dictionary-based English to Kannada translation

   (b) Kannada to English translation to form monolingual English Queries
   Table 3 and Figure 3 portrays sample dictionary-based Kannada to English translation and BLEU scores for mixed script queries, respectively.

2. Dictionary based + WE translation

   (a) English to Kannada translation to form monolingual Kannada Queries

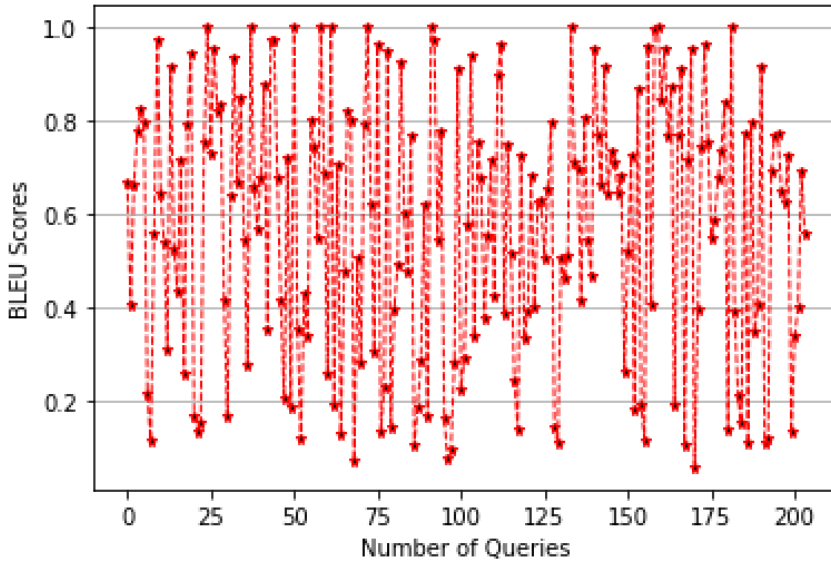

Figure 2. BLEU scores for English to Kannada dictionary translation

Kannada English Input Mixed Script Query	**Translation in English**	**Translations (dictionary)**	**BLEU**	**METEOR score**
ವಿರುಷ್ಕಾ grand reception	Virushka grand reception	ವಿರುಷ್ಕಾ grand reception	0.57	0.62
where ಕಬ್ಬಿಣದ ಕಂಬ located in india	where is iron pillar located in India	where is iron plated pillared located in India	0.65	0.97
earth ಒಳ ಪದರ how much	earth inner layers how much	earth ಒಳ layers how much	0.63	0.75

Table 3. Sample dictionary-based Kannada to English translation

An illustration of results obtained for dictionary-based + WE English to Kannada translation and BLEU scores for mixed script queries are shown in Table 4 and Figure 4.

(b) Kannada to English translation to form monolingual English Queries

An illustration of results obtained for dictionary-based + WE English to Kannada translation and BLEU scores for mixed script queries are shown in Table 5 and Figure 5.

3. Dictionary based + WE + Google translate

(a) English to Kannada translation to form monolingual Kannada Queries

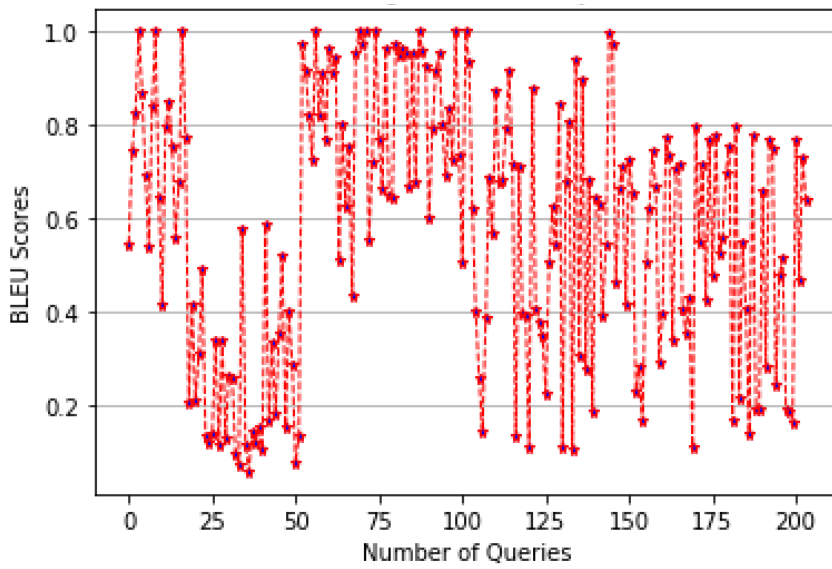

Figure 3. BLEU scores for Kannada to English dictionary translation

Kannada English Input Mixed Script Query	**Translation in Kannada**	**Translations**	**BLEU**	**METEOR score**
ವಿರುಷ್ಕಾ grand reception	ವಿರುಷ್ಕಾ ಅದ್ದೂರಿ ಆರತಕ್ಷತೆ	ವಿರುಷ್ಕಾ ಮಹೋನ್ನತ ಆರತಕ್ಷತೆ	0.81	0.63
where ಕಬ್ಬಿಣದ ಕಂಬ located in india	ಕಬ್ಬಿಣದ ಕಂಬ ಭಾರತದಲ್ಲಿ ಎಲ್ಲಿದೆ	ಎಲ್ಲಿ ಕಬ್ಬಿಣದ ಕಂಬ ಭಾರತದಲ್ಲಿದೆ	0.94	0.67
ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ live telecast	ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ ನೇರ ಪ್ರಸಾರ	ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ live telecast	0.57	0.59

Table 4. Sample Dictionary based + WE English to Kannada translation

Kannada English Input Mixed Script Query	**Translation in English**	**Translations**	**BLEU**	**METEOR score**
ವಿರುಷ್ಕಾ grand reception	Virushka grand reception	ವಿರುಷ್ಕಾ grand reception	0.57	0.62
where ಕಬ್ಬಿಣದ ಕಂಬ located in india	where is iron pillar located in India	where is iron plated pillared located in India	0.75	0.97
earth ಒಳ ಪದರ how much	earth inner layers how much	earth ಒಳ layers how much	0.63	0.75

Table 5. Sample Dictionary based + WE Kannada to English translation

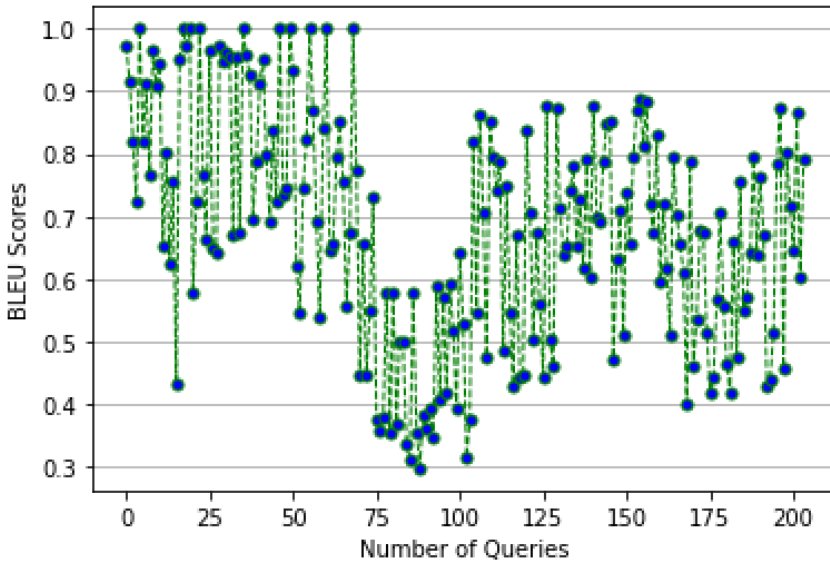

Figure 4. BLEU scores for dictionary based + WE English to Kannada translation

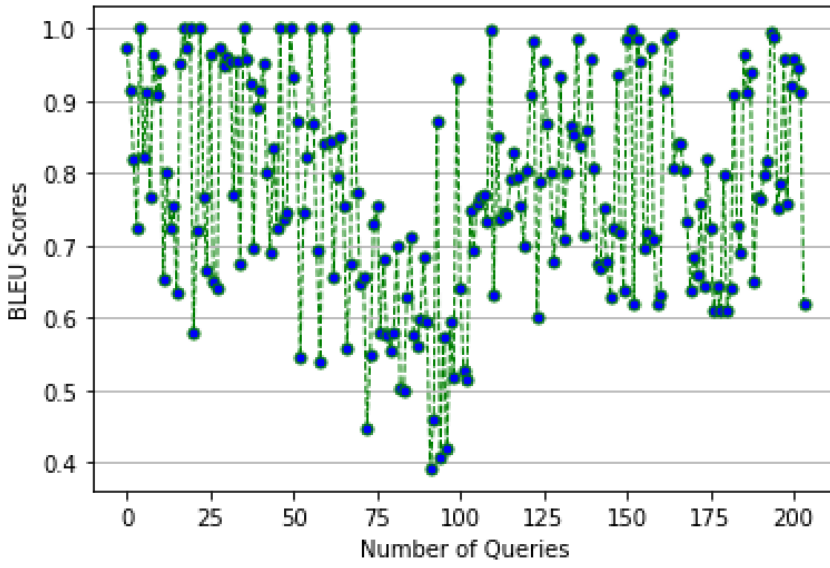

Figure 5. BLEU scores for dictionary based + WE Kannada to English translation

It was observed that translation results were improved by appending google search along with dictionary and WE which is presented in Table 6 and Figure 6.

Kannada English Input Mixed Script Query	**Translation in Kannada**	**Translations**	**BLEU**	**METEOR score**
ವಿರುಷ್ಕಾ grand reception	ವಿರುಷ್ಕಾ ಅದ್ದೂರಿ ಆರತಕ್ಷತೆ	ವಿರುಷ್ಕಾ ಮಹೋನ್ನತ ಆರತಕ್ಷತೆ	0.81	0.63
where ಕಬ್ಬಿಣದ ಕಂಬ located in india	ಕಬ್ಬಿಣದ ಕಂಬ ಭಾರತದಲ್ಲಿ ಎಲ್ಲಿದೆ	ಎಲ್ಲಿ ಕಬ್ಬಿಣದ ಕಂಬ ಭಾರತದಲ್ಲಿದೆ	0.94	0.67
ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ live telecast	ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ ನೇರ ಪ್ರಸಾರ	ರಗ್ಬಿ ವಿಶ್ವ ಕಪ್ ನೇರ ಪ್ರಸಾರ	1.0	1.0

Table 6. Dictionary based + WE + Google translate English to Kannada translation

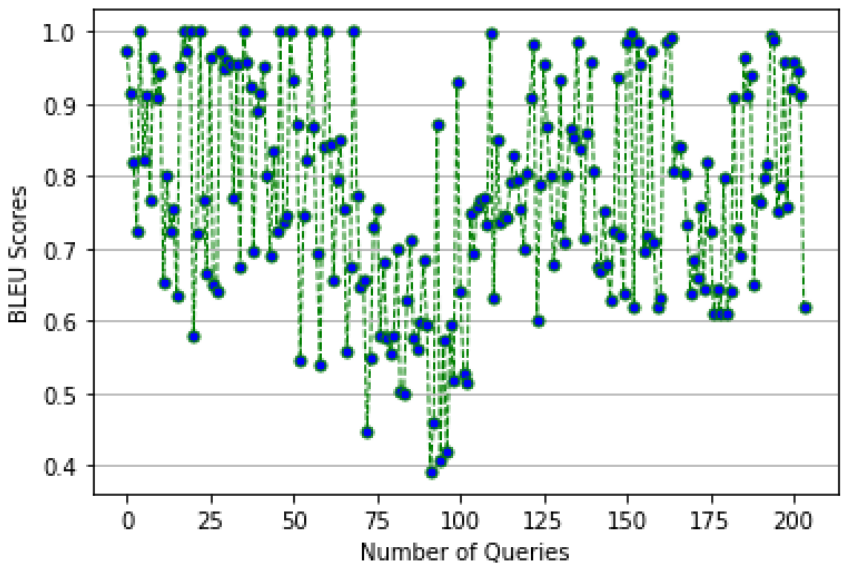

Figure 6. BLEU scores for dictionary based + WE + Google translate English to Kannada translation

(b) Kannada to English translation to form monolingual English Queries
Table 7 and Figure 7 portrays sample dictionary-based + WE + Google translate English to Kannada translation and BLEU scores for mixed script queries respectively.

Kannada English Input Mixed Script Query	**Translation in English**	**Translations**	**BLEU**	**METEOR score**
ವಿರುಷ್ಕಾ grand reception	Virushka grand reception	Virushka grand reception	1.0	1.0
where ಕಬ್ಬಿಣದ ಕಂಬ located in india	where is iron pillar located in India	where is iron plated pillared located in India	0.65	0.97
earth ಒಳ ಪದರ how much	earth inner layers how much	earth inner layers how much	1.0	1.0

Table 7. Dictionary based + WE + Google translate Kannada to English translation

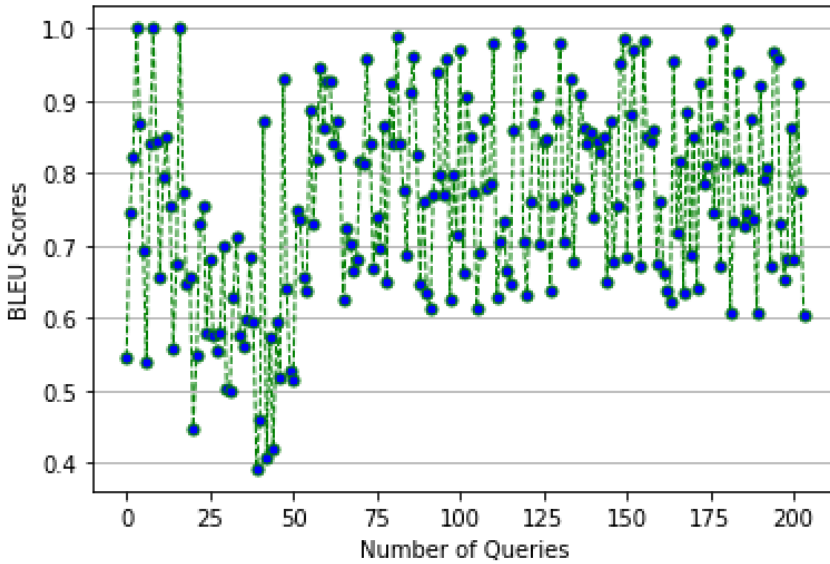

Figure 7. BLEU scores for dictionary based + WE + Google translate Kannada to English translation

It was observed that combination of all three methods, i.e. Dictionary based + word embedding+Google translate, yielded good performance in English to Kannada translation and vice versa. Hence, the method was followed to achieve translations. Words which were not translated by Dictionary based + word embedding + Google translate method were assumed to be NEs and they were transliterated.

5 CONCLUSION

Even though MSIR is a very notable and pervasive problem, it has gained very little attention. In this study, the problem of MSIR is handled for Queries of Kannada

English language pair. A promising solution to address the principal issue of MSIR, i.e., script variations in query was proposed. The MSIR model understands POS of the query terms using BiLSTM-CRF algorithms such that input query words were translated to other language words appropriately. Bilingual dictionary of Kannada and English language was built using Wikipedia dumps to aid translation. An attempt to translate mixed script queries of Kannada and English languages into monolingual queries was done. Three approaches for translation was proposed by constructing bilingual dictionary, word embeddings and Google translate. Proposed approaches were evaluated using BLEU and METEOR metrics. Experimental results shows that proposed Dictionary based + WE + Google translate model achieve better translations than other two models.

Future work includes refinement of the machine translation approach by exploring alternative techniques. One of the refinements could be to make the choice of NMT. As for alternative evaluation techniques, it would be interesting to experiment with other metrics like Translation Error Rate (TER), NIST. Future effort in evaluation would be directed toward character-based metrics which might show the highest correlation with human judgement.

Acknowledgement

The authors of this article gratefully thank the Visvesvaraya Technological University, Jnana Sangama, Belagavi for financial support extended to this research work.

REFERENCES

[1] Yan, Q.—Grefenstette, G.—Evans, D. A.: Automatic Transliteration for Japanese-to-English Text Retrieval. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003, pp. 353–360, doi: 10.1145/860435.860499.

[2] Ahmed, U. Z.—Bali, K.—Choudhury, M.—Sowmya, V. B.: Challenges in Designing Input Method Editors for Indian Languages: The Role of Word-Origin and Context. Proceedings of the Workshop on Advances in Text Input Methods (WTIM 2011), Chiang Mai, Thailand, 2011, pp. 1–9.

[3] Pal, D.—Majumder, P.—Mitra, M.—Mitra, S.—Sen, A.: Issues in Searching for Indian Language Web Content. Proceedings of the 2nd ACM Workshop on Improving Non English Web Searching (iNEWS '08), 2008, pp. 93–96, doi: 10.1145/1460027.1460044.

[4] Dua, N.—Gupta, K.—Choudhury, M.—Bali, K.: Query Completion Without Query Logs for Song Search. Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11), 2011, pp. 31–32, doi: 10.1145/1963192.1963209.

[5] Gupta, K.—Choudhury, M.—Bali, K.: Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics. Proceedings of the Eighth International Conference

on Language Resources and Evaluation (LREC '12), Istanbul, Turkey, 2012, pp. 2459–2465.

[6] Sowmya, V. B.—Varma, V.: Transliteration Based Text Input Methods for Telugu. In: Li, W., Mollá-Aliod, D. (Eds.): Computer Processing of Oriental Languages. Language Technology for the Knowledge-Based Economy (ICCPOL 2009). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 5459, 2009, pp. 122–132, doi: 10.1007/978-3-642-00831-3_12.

[7] Janarthanam, S. C.—Subramaniam, S.—Nallasamy, U.: Named Entity Transliteration for Cross-Language Information Retrieval Using Compressed Word Format Mapping Algorithm. Proceedings of the 2nd ACM Workshop on Improving Non English Web Searching (iNEWS '08), 2008, pp. 33–38, doi: 10.1145/1460027.1460033.

[8] Oard, D. W.—Levow, G.-A.—Cabezas, C. I.: CLEF Experiments at Maryland: Statistical Stemming and Backoff Translation. In: Peters, C. (Ed.): Cross-Language Information Retrieval and Evaluation (CLEF 2000). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 2069, 2001, pp. 176–187, doi: 10.1007/3-540-44645-1_17.

[9] Srivastava, R.—Bhat, R. A.: Transliteration Systems Across Indian Languages Using Parallel Corpora. Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27), Taipei, Taiwan, 2013, pp. 390–398.

[10] Gupta, P.—Bali, K.—Banchs, R. E.—Choudhury, M.—Rosso, P.: Query Expansion for Mixed-Script Information Retrieval. Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14), 2014, pp. 677–686, doi: 10.1145/2600428.2609622.

[11] Pathak, A. K.—Acharya, P.—Kaur, D.—Balabantaray, R. C.: Automatic Parallel Corpus Creation for Hindi-English News Translation Task. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, IEEE, 2018, pp. 1069–1075, doi: 10.1109/ICACCI.2018.8554461.

[12] Li, X.—Zhang, J.—Zong, C.: One Sentence One Model for Neural Machine Translation. 2016, arXiv: 1609.06490.

[13] Sepesy Maučec, M.—Donaj, G.: Machine Translation and the Evaluation of Its Quality. In: Sadollah, A., Sinha, T. S. (Eds.): Recent Trends in Computational Intelligence. IntechOpen, 2019, doi: 10.5772/intechopen.89063.

[14] Dunder, I.—Seljan, S.—Pavlovski, M.: Automatic Machine Translation of Poetry and a Low-Resource Language Pair. 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatia, Croatia, IEEE, 2020, pp. 1034–1039, doi: 10.23919/MIPRO48935.2020.9245342.

[15] Khan, S.—Usman, I.: A Model for English to Urdu and Hindi Machine Translation System Using Translation Rules and Artificial Neural Network. The International Arab Journal of Information Technology, Vol. 16, 2019, No. 1, pp. 125–131.

[16] Gadde, P.—Yeleti, M. V.: Improving Statistical POS Tagging Using Linguistic Feature for Hindi and Telugu. International Conference on Natural Language Processing (ICON-2008), 2008.

[17] Ekbal, A.—Bandyopadhyay, S.: Part of Speech Tagging in Bengali Using Support Vector Machine. International Conference on Information Technology (ICIT '08), Bhubaneswar, India, IEEE, 2008, pp. 106–111, doi: 10.1109/ICIT.2008.12.

[18] Devadath, V. V.—Sharma, D. M.: Significance of an Accurate Sandhi-Splitter in Shallow Parsing of Dravidian Languages. Proceedings of the ACL 2016 Student Research Workshop, Berlin, Germany, ACL, 2016, pp. 37–42, doi: 10.18653/v1/p16-3006.

[19] Antony, P. J.—Mohan, S. P.—Soman, K. P.: SVM Based Part of Speech Tagger for Malayalam. 2010 International Conference on Recent Trends in Information, Telecommunication and Computing (ITC), Kerala, India, IEEE, 2010, pp. 339–341, doi: 10.1109/itc.2010.86.

[20] Ganesh, J.—Parthasarathi, R.—Geetha, T. V.—Balaji, J.: Pattern Based Bootstrapping Technique for Tamil POS Tagging. In: Prasath, R., O'Reilly, P., Kathirvalavakumar, T. (Eds.): Mining Intelligence and Knowledge Exploration. Springer, Cham, Lecture Notes in Computer Science, Vol. 8891, 2014, pp. 256–267, doi: 10.1007/978-3-319-13817-6_25.

[21] Shambhavi, B. R.—Kumar, P. R.: Kannada Part-of-Speech Tagging with Probabilistic Classifiers. International Journal of Computer Applications, Vol. 48, 2012, No. 17, pp. 26–30, doi: 10.5120/7442-0452.

[22] Graves, A.—Schmidhuber, J.: Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. Neural Networks, Vol. 18, 2005, No. 5-6, pp. 602–610, doi: 10.1016/j.neunet.2005.06.042.

[23] Liu, X.—Zhao, J.—Sun, S.—Liu, H.—Yang, H.: Variational Multimodal Machine Translation with Underlying Semantic Alignment. Information Fusion, Vol. 69, 2021, pp. 73–80, doi: 10.1016/j.inffus.2020.11.011.

[24] Laville, M.—Hazem, A.—Morin, E.: TALN/LS2N Participation at the BUCC Shared Task: Bilingual Dictionary Induction from Comparable Corpora. Proceedings of the 13th Workshop on Building and Using Comparable Corpora, Marseille, France, 2020, pp. 56–60.

[25] Sowmya Lakshmi, B. S.—Shambhavi, B. R.: Automatic English to Kannada Back-Transliteration Using Combination-Based Approach. In: Sridhar, V., Padma, M., Rao, K. (Eds.): Emerging Research in Electronics, Computer Science and Technology. Springer, Singapore, Lecture Notes in Electrical Engineering, Vol. 545, 2019, pp. 159–170, doi: 10.1007/978-981-13-5802-9_15.

B. S. Sowmya Lakshmi received B.E. degree from the Visvesvaraya Technological University (VTU) in 2011 and M.Tech. in 2013. In 2021 she completed her Ph.D. from VTU in the field of natural language processing and information retrieval. She has academic experience of about 2 years and published more than 10 research papers in international journals and conferences. Currently she is Assistant Professor in the Department of Machine Learning, BMSCE, Bangalore.

B. R. Shambhavi completed her Ph.D. from the Visvesvaraya Technological University in the area of natural language processing. She has academic and industry field experience of about 13 years. Her areas of interest are natural language processing and information retrieval. She has published more than 20 research papers in international journals and conferences. Currently she is Associate Professor in the Department of ISE, BMSCE, Bangalore. She is a life member of Indian Society for Technical Education (ISTE).