ISSN 2278-3091
Volume 9 No.2, March - April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse214922020.pdf
https://doi.org/10.30534/ijatcse/2020/214922020

A Framework for English-Odia Cross-Language Information Retrieval System

Gouranga Charan Jena, Siddharth Swarup Rautaray
School of Computer Engineering,
KIIT University, Bhubaneswar. India.
{jenagouranga2000, sr.rgpv}@gmail.com

ABSTRACT

This paper depicts a framework for English-Odia Cross-Lingual Information Retrieval system. The system retrieves Odia documents in response to query given in English or Odia. Thus monolingual and cross-lingual information retrieval can be achieved by using this system. Odia is the prominent regional language of Odisha and the sixth classical language of India. It is spoken by more than 33 million people in Odisha and is the official language of Jharkhand state in India. Here we have used an online bilingual dictionary for query translation. This bi-lingual dictionary contains sixteen thousand words including noun, verb, adjective, and adverb. We are using a bilingual dictionary for query translation. Other linguistic resources like tokenizer, stemmer and stop word list etc. for Odia were developed during this work.

Key words: Cross-Lingual Information Retrieval, Vector Space Model, Odia, Document ranking, Bilingual dictionary.

1.INTRODUCTION

The main goal of Information retrieval (IR) is to retrieve all relevant documents (recall) before non-relevant documents, where the user's information need is formally represented in the form of set of free text words known as queries. Another most important feature of Information retrieval system is document indexing, query analysis and query evaluation. The efficiency of an Information retrieval (IR) depends on two performance measures i.e. Precession and Recall. Precession means to retrieve most relevant documents and Recall means to retrieve all relevant documents from the search space. Cross-Lingual Information Retrieval (CLIR) is information retrieval technique by which a user fires a query in one language (L) and retrieves the results in another language (L`). How the users retrieve one or many relevant documents in a language other than the query language? It is only possible if the user's query can be translated to the target language and then fired to the search system for document retrieval known as post-processing. . Due to this additional translation step, it may cause a reduction in the retrieval performance of the CLIR system as compare to monolingual IR. There are many drawbacks in this query translation approach [1] such as missing vocabulary in the dictionary, missing general terms

and wrong translation due to ambiguity [2]. For this we built bi-lingual dictionary contains sixteen thousand words including noun, verb, adjective, and adverb. The idea is to translate each term in the query with an appropriate term from the dictionary. This proposed framework can play a dual role i.e. Monolingual and Cross-Lingual document retrieval system. As we aware that internet becomes pervasive and off course CLIR helps the users to break the barrier of languages to access valuable information in different languages. Odia belongs to the family of Indo-Aryan Languages. It is 6th classical language of India. Now the volume of Odia electronic data has increased very much in World Wide Web. Many Odia new papers, magazines, Govt. websites has seen in Odia language. Mainly for this reason CLIR is a hotspot research field now not only in India but in the whole world.

In this paper we are building an English-Odia CLIR system using vector space model (SVM). It's a widely used technique in Information retrieval field. By this model the user retrieves relevant documents that are similar to the user's input query from a given set of documents. Here the null hypothesis is that the existing search engines don't have the complete capability to retrieve the relevant documents in other languages. The most important features of SVM is it compares the terms in the document collection and then compute the similarity between each document in the collection based on the query terms. Then it sorts the documents in decreasing order of the similarity scores. This model then displays a ranked list of the documents to the user. The top ones are more relevant to the user query as judged by the system. It ranks the documents according to the similarity metric (i.e. cosine) between the

query and document. The smaller the angle between the document and query the more similar they are believe to be. Documents are represented by a term vector and queries are represented by a similar vector. Different Adhoc weighting (term frequency * inverse document frequency) schemes are used.

The rest of the paper we discuss about the CLIR techniques, literature survey, proposed framework along with details description of the SVM model and evaluation results.

2. RELATED WORKS

CLIR development for Indian languages is an emerging research field now. And also it is an important research field due to the existence of many languages used by people across the globe. CLIR provides a new approach in searching document through multitude varieties of languages across the globe. Most of the CLIR research carried out today in Indian languages, involved only few famous languages like Hindi, Tamil, Bengali, Telugu, and Malayalam etc. If you think about the world, then it would be English, French, Chinese, Spanish, and Japanese etc.

There is no significance difference between monolingual and cross-lingual IR if you look into the architectural point of view like post processing, retrieval model and indexing algorithm. The only difference is translation i.e. query translation or document manually or automatically. Again the query translation can performed in various approaches such as dictionary-based machine translation, parallel corpora based statistical lexicon and ontology-based methods.
Typically, the CLIR techniques are classified in to two categories depicted in diagram Figure 1.[3]

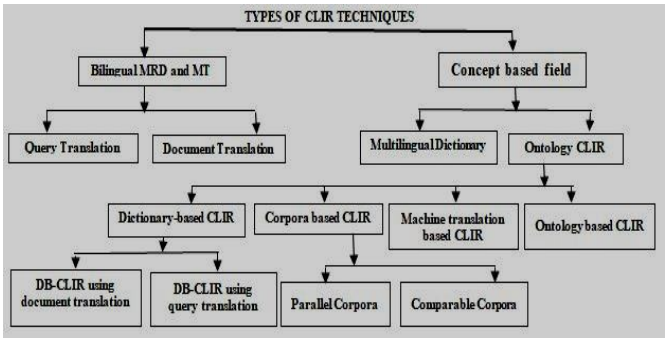

Figure 1: CLIR techniques

Through the Table 1 we intended to provide a crisp detail of various CLIR research had done on Indian languages in past decades.

Table 1: CLIR Techniques for Indian Languages

Languages	Translation Approaches
English to Hindi Larkey & Connell (2003)	Probabilistic dictionary derived from parallel corpus
Hindi to telugu to English P.Pingali & V.Verma (2006)	Bilingual Dictionary
English to Hindi A.Seetha , S.Das & M. Kumar (2007)	Select first equivalent/ preferred –n/ random nth equivalent/ all equivalents from Bilingual dictionary
Bengali & Hindi to English D. Mandal & P. Banerjee (2007)	Machine Translation using Bilingual dictionary
English to Hindi & Hindi to EnglishS. Sethuramalingam & V. Varma (2008)	Bilingual Dictionary
Hindi to English R. Udupa & J. Jagarlamudi (2008)	Bilingual dictionary developed in house
Tamil to English S. Saraswathi & A. Siddhiqaa (2010)	Machine translation and Ontological tree
English to Hindi A. Seetha, S. Das, J. Rana & M. Kumar (2010)	Translation by Shabdanjali dictionary & query expansion by Hindi Wordnet.
Tamil to English	Bilingual dictionary
Tamil to English D.Thenmozhi & C. Aravindan (2010)	Machine Translation
Tamil to English Pattabhi R.K Rao and Sobha. L (2010)	Bilingual dictionary and ontology
English to Bangla A.Imam & S. Chowdhury (2011)	SMT using parallel corpus

3. PROPOSED FRAMEWORK

The major objective of this work is to develop a new framework for Cross language information retrieval using bilingual machine readable dictionary. From the literature discussed in the Table-1 it is clear that the translation approach is a distinct aspect that contributes to the success of the CLIR system. We have implemented this proposed framework using one of the popular IR model i.e. Vector Space Model (SVM). This model has a couple of advantages:

i) It is an algebraic model based on linear algebra.
ii) Term weights are not binary like in Boolean model.
iii) It computes degree of similarity between queries and documents.
iv) Ranking and retrieval of documents is performed according to the similarity measure in decreasing order.

The SVM model not only useful for CLIR system but can be useful in other information retrieval related applications, such as topic tracking, text categorization, and information filtering. Figure 2 shows the overall process of proposed CLIR system.

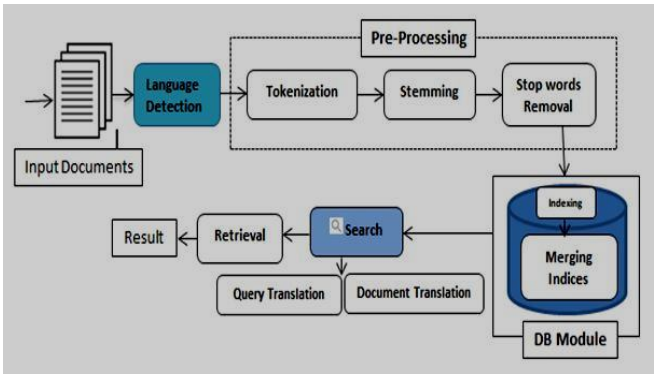

Figure 2: System Architecture of English-Odia CLIR system

The overall paradigm is to have a modular and lightweight framework that can be implemented English-Odia CLIR as shown in Figure 2.

3.1 Vector Space Model (SVM)

In SVM model, first we represent the text documents and query into vector of words and in second step we transform to numerical format so that we can create term-document matrix after the preprocessing. Preprocessing step involves:
Query preprocessing, Document preprocessing and Vector creator.

The detailed workflow of each component is explained below.

a) Query Preprocessor

In the query pre-processing stage the following components are involved in user query:

i. Tokenizer
ii. Stemmer
iii. Stop word Removal

The detail description of the Tokenizer and Stemmer is present in our previous research [4] with example. When the user fires a query to the proposed system in English or Odia then it goes through different stages like tokenizing, stemming and stop word removal. The output is a bag of weighted query words.
Proper nouns and nouns weigh the highest. A bilingual English-Odia lexicon has been used for English query translation to its corresponding Odia words. Odia stemmer developed was used for Odia. Porter stemmer used for English. The unavailability of any query terms in bilingual lexicon is considered as OOV (Out-of-Vocabulary) terms.

b) Document Preprocessor

In the document preprocessing stage, each document in the corpus passes through the different stages like tokenizing, stemming and stop word removal. It is ultimately reducing the number of words to be stored. In the indexer step, these words are indexed using Lucene engine and stored in a hash table called the inverted index file .This file contains an index of term and a postings list for each term. The postings list contains the documents id and the frequency of occurrence of the term in that document. The Count Field in the postings list contains the total number of times that term appear in all the documents of the collection and the Document Frequency is the number of documents containing that term. The structure of inverted index file and postings list is shown in the below figure.

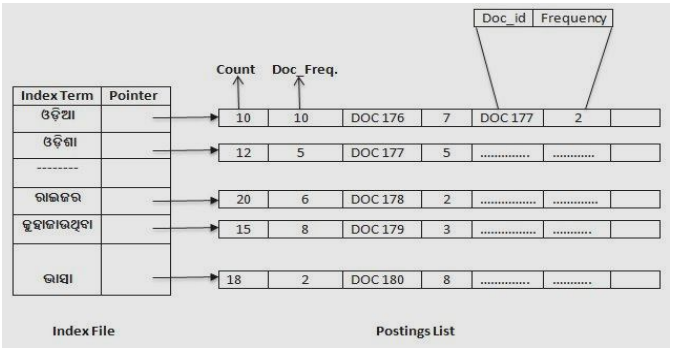

Figure 3: Structure of Inverted Index file and postings list

c) Vector Creator and Processor

In this research we are building an English-Odia CLIR engine or document retrieval system using Vector Space Model (SVM). This approach was used for document ranking and retrieval. This approach widely used in information retrieval systems. Each document is represented as a set of terms represented as document vector. A union of all these set of terms, each set representing a document, forms the 'document Space' of the Corpus. Each distinct term in the union set, represents one dimension in the document space. These terms are the words remaining after operations like stop word removal and stemming. These terms are otherwise known as root/stem to the respective language. Proper noun term weighting improves the performance of the CLIR system using this model.

We can assign a numeric weight to each term in a given document, representing an estimate of the usefulness of the given term as a descriptor of the given document. It is also important to calculate the term weightings because we need to find out terms which uniquely define a document. We should note that a term which occurs in most of the documents might not contribute to represent the document relevance whereas less frequently occurred terms might define document relevance. If a term receive a different weights defines it may be a better descriptor of that document than another else if the term having zero weights means is not present in the document. The weights assigned to the terms in a given document D1 can then be interpreted as the coordinates of D1 in the Document Space. A weighting scheme is composed of three di□erent types of term weighting: local, global, and normalization. The term weight is given by $X_{i,j}$ Y_i N_j where $X_{i,j}$ is the local weight implies a functions of how many times

each term i appears in a document j, Y_i is the global weight implies a functions of how many times each term i appears in the entire collection and N_j is the normalization factor for document j and compensates for discrepancies in the lengths of the documents. It varies from one document to another. This formula depends only on the frequencies within the document and they do not depend on inter-document frequencies . In this case a term weighting scheme for the query terms and a modified form of augmented normalized term frequency for the terms in the document. The below equation employs for calculating the term weighting [5].

$$tf_{i,j} = C * \chi(k_i, A_d) + (1-C) * \frac{\log(freq(k_i, A_d))}{\log(\max\{freq(k_1, A_d)..., freq(k_t, A_d)\})} \quad (1)$$

Where $\chi(k_i, A_d) = \begin{cases} 1 \text{ if that term is present} \\ 0 \text{ if not present,} \end{cases}$

C is a constant set as 0.3 , freq(k_i,A_d) is the term frequency in the document, and (max{freq(k_1,A_d)⋯freq(k_i,A_d)}) is the maximum term frequency in any document. Y_i gives a 'discrimination value' to each term. We remove stop words from documents so that we have fewer global term weighting to handle. Inverse Document Frequency (idf) [6] equals to 0, if a given term present in every document in the corpus. That means less frequent term appear in the corpus has more discriminating. We have used a variation of idf called Probabilistic Inverse Document Frequency (pidf) for calculating Y_i. It assigns weights ranging from for a term that appears in every document (0) to log(n−1) for a term that appears in unique document. It di□ers from idf because probabilistic inverse actually awards zero weight (in practice) for terms appearing in more than half of the documents in the collection.

$$pidf_i = log\frac{N - df_t}{df_t} \quad \text{-------- (2)}$$

Normalization factor (N_i) is useful to correct discrepancies in document lengths and to normalize the document vectors so that documents are retrieved independent of their lengths. Due to lack of N_i, short documents may not be recognized as relevant. The main reasons behind the use of normalization in term weights are as follows:

i) Basically in a long document the same term repeats, that's why the term frequency is higher.

ii) Number of distinct is higher in long document over shorter document. It ultimately gives more preference to long document over short document based on query terms match in the long document in the retrieval process.

Cosine similarity [7] is used to retrieve relevant documents from a user's query we have to calculate cosine angle between each document vector and the query vector to find its closeness. We have also calculated the similarity score between each document vector and the query term vector by applying cosine similarity. Finally, whichever documents having high similarity scores will be retrieved first and considered as relevant documents to the user's query term. It will retrieve ranked documents in decreasing order of their similarity scores. Cosine Normalization is commonly used normalization technique in SVM model. It resolves both the reasons described above for normalization in one step. With an inverted file, the number of postings lists accessed equals the number of query terms. The computational cost is less for shorter queries. The computation factor is expensive Cosine Normalization because of the term

$$\sqrt{\sum_{j=1}^{m} (W_{i,j})^2}$$

Normalization factor allows the system access to all participating document's terms, not only the query terms. To approximate the e□ect of normalization the square root of the number of terms in a document was used as the normalization factor. In the document processing, a log file that contains the document id and the corresponding square root of the number of terms in the document is prepared. This is used at the time of normalization. With this approximation, similarity between documents i and query Q is equation 3.

$$Sim(Q, D_i) = \frac{\sum_{j=1}^{T} (q_j * W_{j,i})}{\sqrt{NumberofTermsinD_i}} \quad \text{--(3)}$$

Where, numerator is the dot product of the query and document i and denominator is the normalization factor used as document length. T is the total no. of keywords in the query Q.

4.USER INTERFACE

We have not fully integrated all the modules in the proposed English-Odia CLIR system yet. We have tested the SVM model by giving some testing data using R language. It is giving the appropriate relevant document to a user query. Below is the snap of the CLIR interface.

Figure 4: User Interface

5.EVALUATION

We have collected the documents and query dataset [8] from FIRE-2018.We are in the process to evaluate [9] these quires against the documents and measure the scores. We have taken 10 documents as shown in below fig. 5.

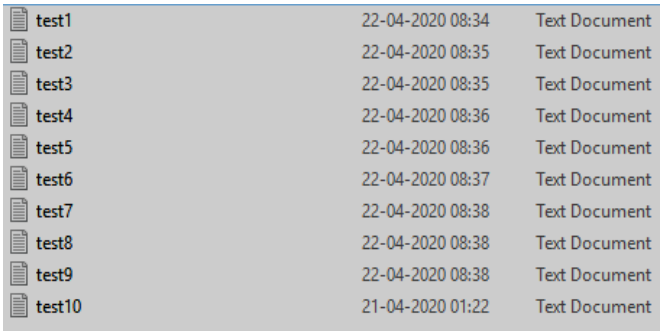

Figure 5: Data Collection

There are 10 English queries (fig-7) have been prepared and given to the CLIR system one by one after translating to Odia (fig-7) for testing.

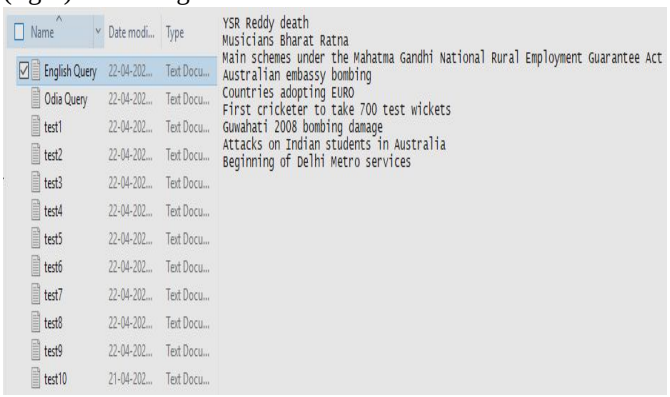

Figure 6: A set of English queries

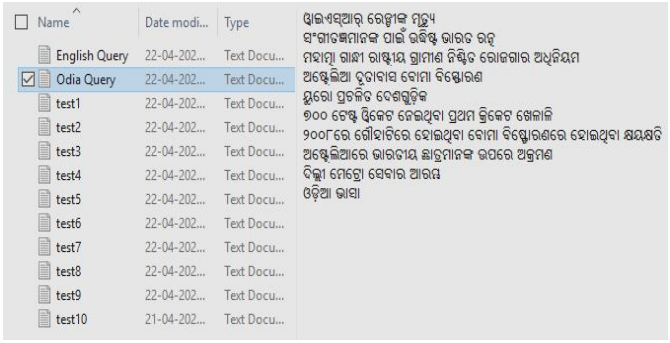

Figure 7: A set of Odia queries

The query retrieves the relevant document as shown in the below figure. The query '୨୦୦୮ରେ ଗୌହାଟିରେ ହୋଇଥିବା ବୋମା ବିସ୍ଫୋରଣରେ ହୋଇଥିବା କ୍ଷୟକ୍ଷତି' retrieves two documents (test7.txt and test4.txt) from the search space (from the ten documents) based on the relevancy. The similarity score is higher in test7.txt (0.44) as compare to test4.txt (0.09) in fig-8.

```
> #test the function
> showTopresults('୨୦୦୮ରେ ଗୌହାଟିରେ ହୋଇଥିବା ବୋମା ବିସ୍ଫୋରଣରେ ହୋଇଥିବା କ୍ଷୟକ୍ଷତି')
       score      docs
10 0.44046540 test7.txt
2  0.36611870 query.txt
7  0.09210737 test4.txt
>
```

Figure 8: Results

6.CONCLUSION

We have developed a framework for English Odia CLIR system. This system based on the query translation approach. We believe that we showed that Cross-Lingual IR framework for English-Odia is viable for a large document collection of document. In future we will try to implement the document translation approach to this framework to improve the performance of this system.

ACKNOWLEDGEMENT

We personally thanks to my friends who directly or indirectly helped me to build the linguistic resources such as stemmer and CLIR interface.

REFERENCES

1. Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerje and Sudeshna Sakar, "Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources," CLEF 2007. Available at http://www.clef-campaign.org/2007/working_notes/ mandalCLEF2007.pdf.
2. Anna R. Diekema., "Translation Events in Cross-Language Information Retrieval," In ACM SIGIR Forum, vol.3,No. 1, June 2004. https://doi.org/10.1145/986278.986296
3. Jena, G., Rautaray, S., "A Comprehensive Survey on Cross-Language Information Retrieval System", Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), 2018 https://doi.org/10.11591/ijeecs.v14.i1.pp127-134
4. Jena, G., Rautaray, S., "Design and Implementation of An Effective Web Based Hybrid Stemmer For Odia Language", International Journal of Engineering and Advanced Technology (IJAAS), 2019. https://doi.org/10.11591/ijaas.v9.i1.pp12-19
5. Nicola Polettini, "The vector Space Model in Information RetrievalTerm Weighting Problem," Department of information and communication Technology, University of Trento, Italy, 2004.
6. Alma'aitah et al.," Document Expansion Method for Digital Resource Objects", International Journal of Advanced Trends in Computer Science and Engineering 8(1.4):17-22,2019. https://doi.org/10.30534/ijatcse/2019/0381.42019
7. W. B. Croft and Raj Das, "Experiments with representation in a document retrieval system," Proceedings of the 13th annual int. ACM SIGIR conf. on

Research and development in information retrieval, Brussels, Belgium, 1989, pp. 349 - 368. https://doi.org/10.1145/96749.98240

8. Balabantaray,R.C.,Sahoo,B.,Swain,M.,Sahoo,D.K.,"IIIT-BH FIRE 2012 Submission: MET Track Odia".
9. Fujii, A., & Ishikawa, T., (2001, March) “Evaluating multi-lingual information retrieval and clustering at ULIS”, In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization. Tokyo. Japan, pp. 5-144.