

Bengali Language Based Bi-Lingual Information Retrieval

Amarinder Kaur

(amarinder.21482@lpu.co.in)

Assistant Professor, School of Computer Science & Engineering
Lovely Professional University, Jalandhar, Punjab

SudhaShanker Prasad

(sudha.21826@lpu.co.in)

Assistant Professor, School of Computer Science &
EngineeringLovely Professional University, Jalandhar, Punjab

Abstract— In this research paper, we have discussed on an experiment that focuses on bi-lingual(Bengali) dictionary-based information retrieval(BLIR) techniques. The Bi-Lingual Information Retrieval(BLIR) involves retrieving of documents based on query in which the query (information required) and/or documents may appear in different Indian languages. It can also be used to enhance the overall user experience to search and retrieve documents in other language in respective of language in which query has been given. We will experiment on the retrieval of English documents in respective of queries in Indian language i.e. Bengali. Our approach is basically focuses on bi-lingual dictionary based retrieval that tries to produce an equivalent English query based on Bengali query. Due to limitation in coverage of bi-lingual dictionary in Bengali language. It affects overall retrieval effectiveness. We have described the findings of different query structures type, query translation techniques and linguistic tools used for BLIR.

Keywords— Bi-Lingual, Document Translation, Data Mining, Machine Translation, Information Retrieval

I. INTRODUCTION

In the generation of web 3.0, Number of Cross-Lingual documents that are web-accessible has helped many of us by enabling them to understand documents in multiple language. India has almost 22 official languages, where most of its inhabitants Cross-lingual in nature, and aware to English and multiple native languages, in addition to mother tongue. Bi-lingual Information Retrieval involves raising a query in one language and finding the set of relevant documents against that query in different languages other than language in which query has been given. It majority helps the user by looking, finding relevant and assessing information without being limited by to a particular language. Mostly, Bi-lingual IR helps user by not restricting them in his own particular language but allowing the environment to return set of document in a different language. Bengali Language has very poor/sparse resources, due to this limitation it has very few language system, resources or tools (i.e. morphocreater, stemmer, dictionary, etc.). Various ideas that involves machine translation, Bi-Lingual dictionary, Parallel corpora and have been tried to address this problem [1][2][3][4]. However, in our research experiment we have tried till Bi-Lingual Dictionary Based Retrieval cum Translation. The rest of the paper is arranged as follows. Section II presents Related Works in BLIR & also including various Indian languages. Section III carries technique on translation. Section IV deals majority towards creation and experiments carried out with our BLIR system. Section V & VI deals with result obtained from our system & discussion respectively. In Section VII, we conclude this paper with Conclusion & Future Works.

II. RELATED WORKS

Bi-Lingual Information Retrieval including Bengali Indian language is entirely a new space of experiment & research among Natural Language Processing group. Some of the old works in this area includes work done during TIDES exercise [5] in 2003. Larkey et al. used Language modelling approach [6] and showed the various ways of stemmer for highly incurvate languages, like Bengali. It was also helpful for Telgu and Bengali to English BLIR by Prasad et al. [7]. BLIR involving Indian languages and their different ideas are also discussed in [8] [9] [10] [11]. Sivaji et al. [12] has tried a zonal indexing approach for Bengali, Telugu and Hindi languages documents. Documents were first created into some divided zones and then re-assigned some unique weightage to the terms, and then the relevant count of each term is then re-calculated based on frequency zones. Thereafter frequency is used as an index for creation query. A word alignment system was created that uses SMT system for learning and training of parallel aligned sentences, i.e. used convert the query into English is proposed by Jagadeeshet al.[13]. Nandkishor and Vasnik's [14], research involves the Natural Language Processing techniques approaches for boosting the quality of the search on different e-resources. Saurabh et al. [15] proposed an approach for improving the performance of the Hindi-English BLIR system. Work is based on finding all available combination of Hindi query achieved using translation using transliteration of English query terms. Herbert et al. [16] in their work proposes query translation obtained by noun

mapping phrases in the query. Chinnakotla et al. [18] presents a Hindi-English & Marathi-English Cross-Lingual IR System that uses dictionary-based query translation. Thenmozhi and Aravindan [17] present a English-Tamil Cross Lingual IR System for Agriculture Group. Mandal et al. [19] presents a Bengali-English & Hindi-English Cross Lingual IR system that is based on Direct Generation of Query and Translation approach using Machine Learning. Hanumanthappa and Reddy[20] present English-Kannada & English-Telgu Cross-Lingual IR System as part of Ad-Hoc Bilingual task.

III. BI-LINGUAL TRANSLATION TECHNIQUES

There are majority two translation techniques to Bi-Lingual Information Retrieval Namely Bi-Lingual Dictionary Based Retrieval cum Translation, Machine based Bi-Lingual Translation. Brief insights of these techniques has been discussed below:

3.1 Bi-Lingual Dictionary Based Retrieval cum Translation –Machine-Understandable Bi-Lingual Dictionaries are accessible and can be used in modules that helps in translation of BLIR engines. A dictionary-dependent translation approach is very simple but it has mainly two limitations such as lack of coverage & ambiguity.

3.2 Machine based Bi-Lingual Translation–Machine based Bi-Lingual Translation tries translation of the text from one language to another. In Bi-lingual IR, Machine Translation is duly accepted due to its wide availability & required linguistic resources to pre-train them. It tries translation of text through a software tool that take input in one language and outputs in another language. Its only performs the one to one translation of words from one language to another.

IV. EXPERIMENTS

In our Bi-Lingual IR system the queries and documents may appear Bengali or English language. In this experiment, we tried retrieval of English document based on Bengali queries that tries Machine based Bi-Lingual Translation approach, viz. tokens, Removal of stop-words & Stemming. The similar English query from Bengali language is created using Query Translation through Bi-Lingual dictionary lookup and Query Transcribe. One of the major issue we faced in our experiments was to handle words not present in dictionary. We transcribed out of dictionary words by using bi-lingual lexicons effectively. We had access to a Bengali-English Bi-Lingual lexicon of ~32K Bengali words.

In order to achieve our task of document based retrieval with the limitation of limited resources, we tried Query Translation followed by Query Transliteration. Normalized scores are used for document retrieval, for the query based on English Query by feeding them into Lucent Search Engine(LSE), which internally uses the VSM of Information Retrieval used for tokens & indexes. Given below Fig.1 depicts flowchart of given approach.

4.1 Translation of Query –

On each word from query, firstly we perform stemming, followed by stemmed terms are looked into machine understandable Bi-Lingual dictionary. If the particular word is found in dictionary, all corresponding translation is used to create a similar English query. During translation of Query, Parts-of-Speech (POS) information from the query words are not taken into care. We noticed that most of the query words are not present in the lexicon possibly due to the limitation in size of the dictionary, improper stemming.

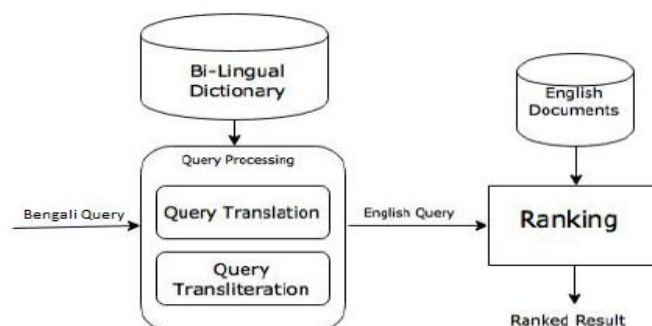


Fig. 1. System Description for Proposed Approach

4.2 Transliteration of Query –

The query words not present in dictionary are transliterated into English using transliteration. The task is carried out on character basis & then transliterated word is created by combining individual character. Due, to various limitation in Bengali & English language such as Phonetics, we actually ended up having more than a possibilities of a word. For example, the Bengali

term for Patna may have multiple transliterations possible as output: Patna, Paatna and Patnaa. To create a unique result, we carried manual matching of the terms in respect of manually-built entity list through Edit Distance algorithm. For e.g., list built manually may correctly returned Patna as the final term in Bi-Lingual runs. Pseudo Relevance Feedback in respect of Query Expansion doesn't reflect a better result in BLIR, but it brings down the overall performance of result as it also considered significant amount of noise.

V. EXPERIMENTS

To determine the performance of Bi-Lingual IR, we created 4 set of document collection in which each 40 Bengali queries results 650 documents. We described our finding for 4 runs in Table 1 in this section. The findings for each of the runs mentioned in Table 1 are: column 1 describes the number of run, rest of the column denotes different metrics such as ret_rel(Relevant Documents Retrieved), preci(Precision), reca(Recall). Each row describes the finding of various runs to their corresponding score against metrics.

Table 1 Experiment Result

RUN_TYPE	Retrieved	Relevant	ret_rel	preci	reca
RUN_1	650	210	124	0.19	0.59
RUN_2	650	210	138	0.21	0.65
RUN_3	650	210	144	0.22	0.68
RUN_4	650	210	132	0.20	0.62

We observed an abnormal finding for Bengali due to lack of vast dictionary in translation of query during query creation. We noticed that queries having more named entities gives a better result in respect of queries having less entities. Increasing the lexical entries, results in addition of more 'noisy words' into resultant query, as we are not using any translation disambiguation algorithm. Hence, resultant performance of our BLIR system is affected.

VI. DISCUSSION

We noticed that lackness in number of better stemmers for Bengali languages is one of crucial reason that resulting aberrant behaviour of our BLIR system. One of major reasons that contributed for hampering the overall result is lackness of the machine-readable dictionaries for Bengali languages. We also noticed, the nos. ambiguities are added as a result of simple suffix removal. With our analysis, we also noticed the need a better stemmer in order to handle inflective Bengali languages. We also noticed English terms found in English-Bengali Bi-lingual dictionary may linked to some different sense of source language word or synonyms. One Major issues, that we noticed is lackness of words in bi-lingual dictionary. Generally, such cases are dealt with phonetic query transliteration.

Due to this, our overall performance of system is majority dependent on transliteration system. We too noticed that use of Sense Word Disambiguation can be a more effective solution for better overall performance of our BLIR system.

VII. CONCLUSION & FUTURE WORK

This paper reports finding of an experiment towards Bengali to English Bi-lingual information retrieval & its evaluation findings & results. We noticed that overall degraded performance our BLIR system. It clearly mentions importance of better & rich bi-lingual dictionaries for Indian languages. Creation of better stemmer for Indian languages plays an important role towards creating a better & efficient BLIR. Our future work will target towards building effective Bi-lingual retrieval system for Indian languages by building a better transliteration system along with better bi-lingual dictionaries. We will also try explore its scope in other Indian languages such as Hindi, Punjabi, etc.

REFERENCES

- [1] Monz, Christof, and Bonnie J. Dorr. "Iterative translation disambiguation for cross-language information retrieval." *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 520-527. ACM, 2005.D. Oard, "The surprise language exercises", *ACM Transactions on Asian Language Information Processing*, vol. 2, no. 2, pp. 79-84, 2003.
- [2] Bertoldi, Nicola, and Marcello Federico. "Statistical models for monolingual and bilingual information retrieval." *Information Retrieval* 7, no. 1-2 (2004): 53-72.
- [3] Hull, David A., and Gregory Grefenstette. "Querying across languages: a dictionary-based approach to multilingual information retrieval." *In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 49-57. ACM, 2016

- [4] Diekema, Anna R. "Translation events in cross-language information retrieval." In *SIGIR FORUM*, vol. 38, no. 1, pp. 75-75. 2004..
- [5] Larkey, Leah S., Margaret E. Connell, and NasreenAbduljaleel. "Hindi CLIR in thirty days." *ACM Transactions on Asian LanguageInformation Processing (TALIP)* 2, no. 2 (2003): 130-142.
- P. Prasad, K.K. Tune, and V. Varma, "Hindi, telugu, oromo, english CLIR evaluation", Evaluation of Multilingual and Multi-modal Information Retrieval, Springer Berlin Heidelberg, pp. 35-42, 2006.
- [6] Pingali, Prasad, Kula Kekeba Tune, and Vasudeva Varma. "Hindi, Telugu, Oromo, English CLIR Evaluation." In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 35-42. Springer Berlin Heidelberg, 2006.
- [7] Bandyopadhyay, Sivaji, TapabrataMondal, Sudip Kumar Naskar, Asif Ekbal, RejwanulHaque, and Srinivasa Rao Godhavarthy."Bengali, Hindi and Telugu to English ad-hoc Bilingual task at CLEF 2007." In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 88-94. Springer Berlin Heidelberg, 2007.
- [8] Pingali, Prasad, JagadeeshJagarlamudi, and Vasudeva Varma. "*WebKhoj: Indian language IR from multiple character encodings.*" In Proceedings of the 15th international conference on World Wide Web, pp. 801-809. ACM, 2006.
- [9] Pingali, Prasad, and Vasudeva Varma. "IIIT Hyderabad at CLEF 2007-Adhoc Indian Language CLIR Task." In *CLEF (Working Notes)*.2007.
- [10] Pingali, Prasad, and Vasudeva Varma. "*Multi-lingual Indexing Support for CLIR using Language Modeling.*" IEEE Data Eng. Bull. 30, no.1 (2007): 70-85.
- [11] Mandal, Debasis, Mayank Gupta, SandipanDandapat, Pratyush Banerjee, and Sudeshna Sarkar. "*Bengali and Hindi to English CLIR evaluation.*" In Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 95-102. Springer Berlin Heidelberg, 2007.
- [12] Vasnik, Nandkishor, ShriyaSahu, and Devshri Roy. "Talash: A Semantic and Context Basedoptimized Hindi Search Engine."InternationalJournal of Computer Science, Engineering and Information Technology (IJCEIT) 2, no. 3 (2012).
- [13] Varshney, Saurabh, and JyotiBajpai. "Improving Performance of English-Hindi Cross Language Information Retrieval Using Transliteration of Query Terms." *arXiv preprint arXiv:1401.3510* (2014).
- [14] Herbert, Benjamin, GyörgySzarvas, and IrynaGurevych. "Combining query translation techniques to improve cross-language information retrieval." In *European Conference on Information Retrieval*, pp. 712-715. Springer Berlin Heidelberg, 2011.
- [15] Thenmozhi, D., and C. Aravindan. "Tamil-English Cross Lingual Information Retrieval System for Agriculture Society." In *ConferencePapers* Conference papers, p. 173. 2009.
- [16] Chinnakotla, Manoj Kumar, Sagar Ranadive, Om P. Damani, and Pushpak Bhattacharyya. "Hindi to English and Marathi to English cross language information retrieval evaluation." In Workshop of the Cross-Language Evaluation Forum for European Languages, pp. 111-118. Springer Berlin Heidelberg, 2007.
- [17] Mandal, Debasis, SandipanDandapat, Mayank Gupta, Pratyush Banerjee, and Sudeshna Sarkar. "Bengali and Hindi to English Cross Language Text Retrieval under Limited Resources." In *CLEF (Working Notes)*. 2007.
- [18] Reddy, Mallamma V., and Dr M. Hanumanthappa. "Kannada and Telugu Native Languages to English Cross Language Information Retrieval." International Journal of Computer Science and Information Technologies 2, no. 5 (2011): 1876-1880.
- [19] Clough, Paul, and Mark Sanderson. "Measuring pseudo relevance feedback &clir." In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 484-485. ACM, 2004.