Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources

Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, Sudeshna Sarkar
IIT Kharagpur, India
{debasis.mandal, pratyushb} @gmail.com
{sandipan, mayank, sudeshna} @cse.iitkgp.ernet.in

Abstract

This paper describes our experiment on two cross-lingual and one monolingual English text retrievals at CLEF[1] in the ad-hoc track. The cross-language task includes the retrieval of English documents in response to queries in two most widely spoken Indian languages, Hindi and Bengali. For our experiment, we had access to a Hindi-English bilingual lexicon, 'Shabdanjali', consisting of approx. 26K Hindi words. But neither we had any effective Bengali-English bilingual lexicon nor any parallel corpora to build the statistical lexicon. Under this limited resources, we mostly depended on our phoneme-based transliterations to generate equivalent English query from Hindi and Bengali topics. We adopted Automatic Query Generation and Machine Translation approach for our experiment. Other language-specific resources included a Bengali morphological analyzer, a Hindi stemmer and a set of 200 Hindi and 273 Bengali stopwords. Lucene framework was used for stemming, indexing, retrieval and scoring of the corpus documents. The CLEF results suggested the need for a rich bilingual lexicon for CLIR involving Indian languages. The best MAP values for Bengali, Hindi and English queries for our experiment were 7.26, 4.77 and 36.49 respectively.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries.; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation.

Keywords

Bengali, Hindi, Transliteration, Cross-language Text Retrieval, CLEF Evaluation.

1 Introduction

Cross-language (or cross-lingual) Information Retrieval (CLIR) involves the study of retrieving the documents in a language other than the query language. Since the language of query and documents to be retrieved are different, either the documents or queries need to be translated in CLIR. But this translation step tends to cause a reduction in the retrieval performance of

[1]Cross Language Evaluation Forum. http://clef-campaign.org

CLIR as compared to monolingual information retrieval. A study in [1] showed that missing specialized vocabulary, missing general terms, wrong translation due to ambiguity and correct identical translation are the four most important factors for the difference in performance for over 70% queries between monolingual and cross-lingual retrievals. This puts the importance on effective translation in CLIR research. Again, the document translation requires a lot of memory and processing capacity than its counterpart and therefore the query translation is more popular in the IR research community involving multiple languages[5].

Oard [7] presents an overview of the Controlled Vocabulary and Free Text retrieval approaches followed in CLIR research within the query translation framework. But the present research in CLIR are mainly concentrated around three approaches: Dictionary based Machine Translation (MT), Parallel Corpora based statistical lexicon and Ontology-based methods. The basic idea in Machine Translation is to replace each term in the query with an appropriate term or a set of terms from the lexicon. In current MT systems the quality of translations is very low and the high quality is achieved only when the application is domain-specific [5]. The Parallel Corpora-based method utilizes the broad repository of multi-lingual corpora to build the statistical lexicon from the simliar training data as of the target collection. Knowledge-based approaches use ontology or thesauri to replace the source language word by all of its target language equivalents. Some of the CLIR models built on these approaches or on their hybrids can be found in [5][6][8][10].

This paper presents two cross-lingual and one English monolingual text retrieval. The cross-language task includes English document retrieval in response to queries in two Indian languages: Hindi and Bengali. Although Hindi is mostly spoken in north India and Bengali in the Eastern India and Bangladesh only, the former is the fifth most widely spoken language in the world and Bengali the seventh. This requires attention on CLIR involving these languages. In this paper, we restrict ourselves to Cross-Language text retrieval applying Machine Translation approach.

The rest of the paper is structured as follows. Section 2 briefly presents some of the works on CLIR involving Indian languages. The next section provides the language specific and open source resources used for our experiment. Section 4 builds our CLIR model on the resources and explains our approach. CLEF evaluations of our results and their discussions are presented in the subsequent section. We conclude this paper with a set of inferences and scope of future works.

2 Related Work

Cross-language retrieval is a budding field in India and the works are still in its primitive state. The first major work involving Hindi occurred during TIDES Surprise Language exercise in a one month period. The objective of the exercise was to retrieve Hindi documents, provided by LDC (Linguistic Data Consortium), in response to English queries. The participants used parallel corpora based approach to build the statistical lexicon [3][4][12]. [4] assigned statistical weightage on query and expansion terms using the training corpora and this improved their cross-lingual results over monolingual runs. [3][9] indicated some of the language-specific obstacles for Indian languages, viz., propritary encodings of much of the web text, lack of availability of parallel corpora, variability in Unicode encoding etc. But all of these works were the reverse of our problem statement for CLEF. The related work of Hindi-English retrieval can be found in [2].

3 Resources used

We used various language specific resources and open source tools for our Cross Language Infomation Retrieval (CLIR) experiments. For the processing of English query and corpus, we used the stop-word list (33 words) and porter stemmer of Lucene framework. For Bengali query, a Bengali-English transliteration (ITRANS) tool[2] [11], a set of Bengali stop-words[3] (273 words),

[2]ITRANS is an encoding standared specifically for Indian languages. It converts the Indian language letters into Roman (English) mostly using its phoneme structure.

[3]The list was provided by Jadavpur University, Kolkata.

an open source Bengali-English bio-chemical lexicon (9k Bengali words) and a Bengali morphological analyzer of moderate performance were used. Hindi language specific resources included a Hindi-English Transliteration tool (wx and ITRANS), a Hindi stop-word list of 200 words, a Hindi-English bilingual lexicon 'Shabdanjali' containing approximately 26K Hindi words and a Hindi Stemmer[4]. We also manually built a named entity list of 1510 entries mainly drawn from the names of countries and cities, abbreviations, companies, medical terms, rivers, seven wonders, global awards, tourist spots, diseases, events of 2002 from wiki etc. Finally, the open source Lucene framework was used for indexing and retrieval of the documents with their corresponding scores.

4 Experimental Model

The objective of Ad-Hoc Bilingual (X2EN) and Monolingual English tasks was to retrieve the relevant documents from English target collection and submit the results in ranked order. The topic sets for these two tasks consist of 50 topics and the participant is asked to retrieve at least 1000 documents from the corpus per query for each of the source languages. Each topic consists of three fields: a brief 'title', almost equivalent to a query provided by the end-user to a search engine; a one-sentence 'description', specifying more accurately what kind of documents the user is looking for from the search and a 'narrative' for relevance judgements, describing what is relevant to the the topic and what is not. Our approach to the problem can be broken into 3 phases: corpus processing, query generation and document retrieval.

4.1 Corpus Processing

The English news corpus of LA Times 2002, provided by CLEF, contained 1,35,153 documents of 433.5 MB size. After removing stop words and stemming the documents, they were indexed using the Lucene indexer to obtain the index terms corresponding to the documents.

4.2 Query Generation

We adopted Automatic Query Generation method to immitate the possible application of CLIR on the web. The language-specific stop-words were first removed from the topics. To remove the most frequent suffixes, we used a morphological analyzer for Bengali, a stemmer for Hindi and the Lucene stemmer for English topics. We considered all possible stems for a single term as no training data was available to pick the most relevant stem. This constitutes the final query for English monolingual run. For Indian languages, the stemmed terms were then looked up in the bilingual lexicon for their translations into English. All the translations for the term were used for the query generation (Structured Query Translation), if the term was found in the lexicon. But many terms did not occur in the lexicon due to its limitation in size or the improper stemming or as the term is a named entity [2]. Those terms were first transliterated into ITRANS and then matched against the named entity list with the help of an approximate string matching algorithm, edit-distance algorithm. The algorithm returns the best match of the term for the pentagram statistics. This produces the final query terms for cross-lingual runs. The queries were constructed from the topics consisting of one or more of the topic fields.

Note that we did not expand the query using the Pseudo Relevance Feedback (PRF). This is due to the fact that it does not improve the retrieval significantly for CLIR, rather hurts by increasing noise [13], or increases queries in which no relevant documents are returned [4].

4.3 Document Retrieval

The query generated in the above phase is fed into Lucene search engine and the documents were retrieved along with their normalized scores. Lucene scorer follows the Vector Space Model (VSM) of Information Retrieval.

[4]'Shabdanjali' and the Hindi stemmer were built by IIIT, Hyderabad.

5 CLEF Evaluation and Discussions

The evaluation document set for the ad-hoc bilingual and monolingual tracks consists of 1,35,153 documents from LA Times 2002. For the 50 topics originally provided by CLEF, there were manually selected 2247 relevant documents which were matched against the retrieved documents of the participants. We provided the set of 50 queries to the system for each run of our experiments. Six official runs were submitted for the Indian langauges to English bilingual retrieval, three for Hindi queries and three for Bengali queries. Three monolingual English runs were also submitted to compare the results between bilingual and monolingul retrievals. The runs were performed using only <title> field, <title+desc> fields and <title+desc+narr> fields per topic for each of these languages. The performance metrics for the nine runs of our experiments are presented in the following tables.

Table 1: Primary Metrics (in %) for the Official Runs.

Lang	Run	MAP	GMAP	B-Pref	P@10
Bengali	<title>	4.98	0.11	5.43	6.60
	<title+desc>	7.26	0.50	10.38	10.20
	<title+desc+narr>	7.19	0.57	11.21	10.80
Hindi	<title>	4.77	0.21	9.95	6.40
	<title+desc>	4.39	0.32	11.58	8.60
	<title+desc+narr>	4.77	0.34	12.02	8.40
English	<title>	30.56	19.51	29.51	37.80
	<title+desc>	36.49	27.34	34.54	46.00
	<title+desc+narr>	36.12	23.51	35.65	45.60

5.1 Discussions

Table 1 presents four basic primary metrics for CLIR, viz., MAP (Mean Average Precision), GMAP (Geometric Mean Average Precision), B-Preference and Precision at 10 retrieved documents (P@10) for all of our official runs. The lower values of the GMAP corresponding to MAP clearly specifies the poor performance of our retrievals in the lower end of the average precision scale. Also, lower values of the MAP for Hindi than the work of [2] clearly suggests the need for query expansion at the source language end. It is evident from the monolingual English and bilingual Bengali runs that adding extra information to query through <title+desc> increases the performance of the system. But adding the <narr> field has not improved the result significantly. This is probably due to the fact that this field was meant for the relevance judgement in the retrieval and we have not made any effort in preventing the retrieval of irrelevant documents in our IR model. This, in turn, has also affected the MAP value for all the runs. However, the improvement in the result for <title+desc> run over <title> run is not significant for Hindi. This is probably due to the fact that using Structured Query Translation (SQT) increased too much noise in the query to compensate the effect of a better lexicon. Also, we used morphological analyzer for bengali rather than stemmer (for hindi) which was suggested by [2] and this may have contributed to the better result for Bengali.

Table 2 shows the results of the topicwise score breakup for the relevant 2247 documents. As seen from the table, number of failed topics (with no relevant retrieval) and topics with MAP $\leq$ 10% gradually decreased with more fields from the topic, thus establishing the fact again

Table 2: Result of Querywise Score breakup.

Language	Run	Failed	MAP≤ 10%	MAP≥ 50%	Recall (in %)
Bengali	<title>	14	22	19	27.06
	<title+desc>	8	13	22	37.87
	<title+desc+narr>	5	14	23	40.32
Hindi	<title>	10	18	20	31.51
	<title+desc>	6	14	18	30.57
	<title+desc+narr>	4	11	17	30.97
English	<title>	0	1	43	72.54
	<title+desc>	0	0	46	78.95
	<title+desc+narr>	0	0	45	78.19

mentioned in the earlier paragraph. Also, the better result for Hindi than Bengali is clearly attributed to its better lexicon. But when it comes to the number of topics with MAP $\geq$ 50%, Bengali clearly outperforms Hindi due to the ńoise factor; mentioned in the previous paragraph. A careful analysis of the queries revealed that the queries with named entities provided better results for all the runs, whereas the queries without named entities performed very poor due to poor bilingual lexicons and thus brininging down the overall performance metrics. This clearly implies the importance of a very good bilingual lexicon and transliteration tool in the CLIR for Indian languages. Recall is a very important performance metric for CLIR specifically for the case when the number of relevant documents is significantly low compared to the target collection (in this case, it is 1.66% only). It is noteworthy that the recall value has improved even for the <title+desc+narr> field compared to other runs and Bengali has again outperformed Hindi due to ńoise factor.

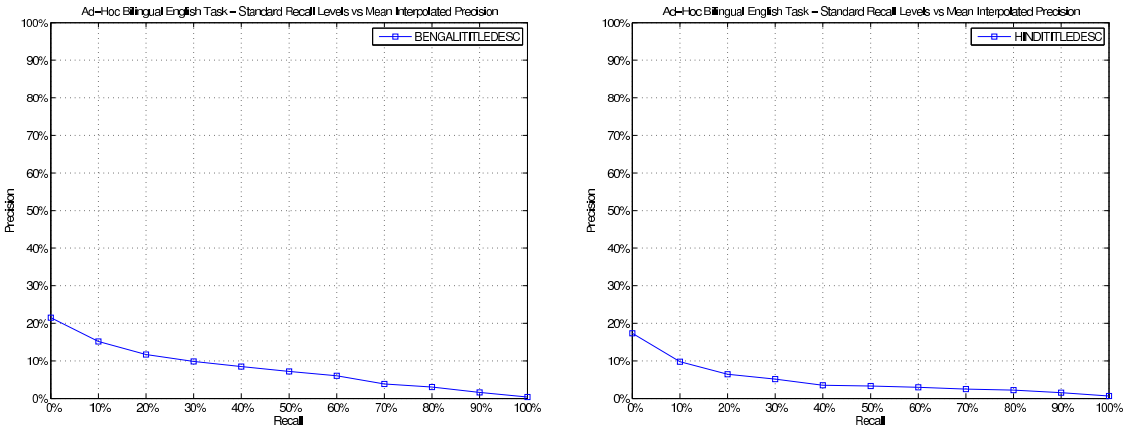

Figure 1: Recall Vs Precision for Bengali and Hindi to English Cross-language runs for <Title+Desc> fields.

The recall vs average precision graphs in Figure 1 and retrieved documents vs precision graphs in Figure 2 suggest the need for refinement of important query terms (e.g. named entity) and weigh them more than the translated terms. Also, we used Structured Query Translation and

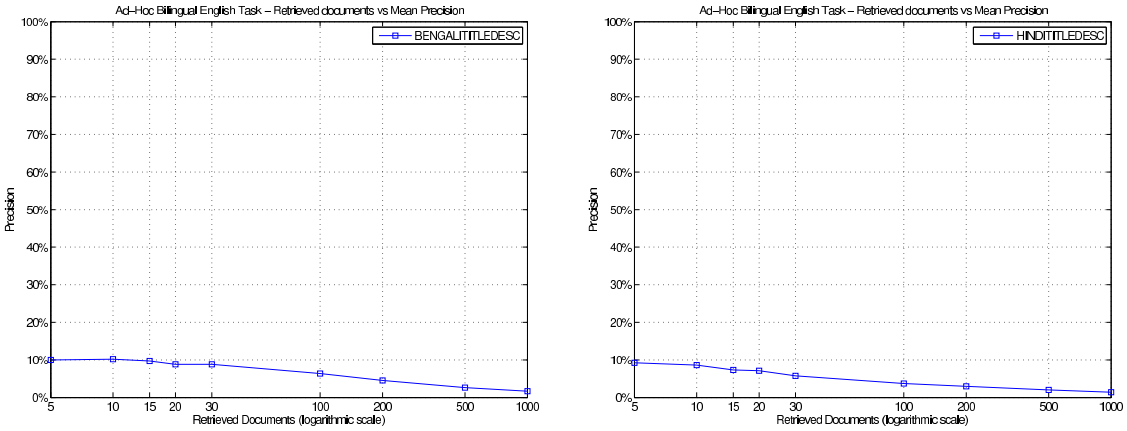

Figure 2: Retrieved documents Vs Precision for Bengali and Hindi to English Cross-language runs for <Title+Desc> fields.

assigned uniform weight on all of them. But this has affected the precision values for some queries even when the recall is significantly high. Again, we used all possible stems for a term when multiple stems are possible and this has also added to the lower precision values for some queries. A proper named entity recognizer is also important to prune out the named entities from other non-lexical terms. All of these will decrease the noise in the final query and thereby help the lucene ranking algorithm to push the relevant documents to the top.

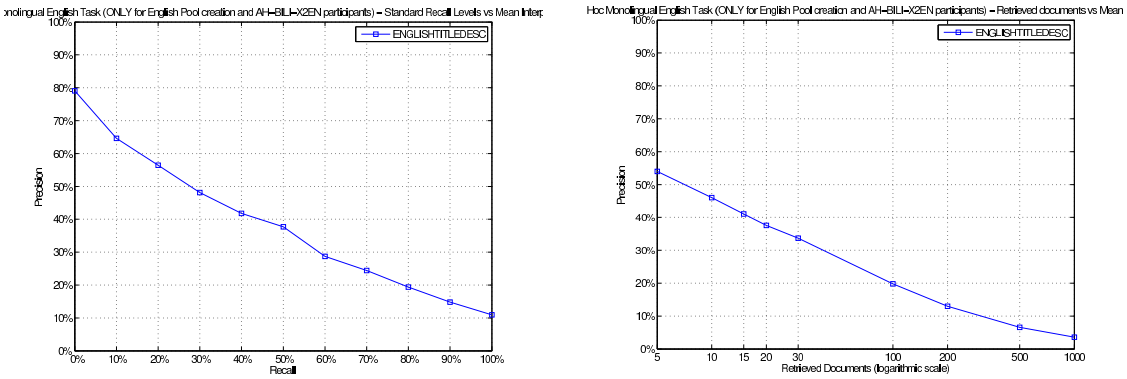

Figure 3: Recall Vs Precision and Retrieved documents Vs Precision monolingual English runs for <Title+Desc> fields.

6 Conclusions and Future Works

This was our first participation in CLEF and we performed our experiment under a limited resource scenario with a very basic Machine Translation approach. But the experiment pointed out the necessity of good language-specific resources, specifically a rich bilingual lexicon. A close analysis between cross-lingual and monolingual retrievals clarly pointed out the importance of four factors in CLIR, mentiond earlier [1]. Apart from the above language-specific requirements, a number of good computational approaches like query expansion by Pseudo Relevance Feedback (PRF) both at the source and target language ends, query refinement by assigning various weightage to query terms, proper stemming, irrelevance judgement to improve ranking, parallel corpus to build the

statistical lexicon, named entity recognizer to prune them out of non-lexical terms, Multi-word Expression (MWE) detection, Word Sense disambiguation to avoid multiple translations (SQT) will also increase the performance of the system. Also, pruning out the irrelevant documents from retrieveing will increase the precision of the results. We will make attempt to experiment and verify their effects in CLIR involving Indian languages in our future works.

References

[1] Anna R. Diekema. Translation Events in Cross-Language Information Retrieval. *In ACM SIGIR Forum*, Vol. 38, No. 1, June 2004.

[2] Prasad Pingali and Vasudeva Varma. Hindi and Telegu to English Cross Language Information Retrieval at CLEF 2006. *In Cross Language Evaluation Forum (CLEF)*, 2006, Spain.

[3] Leah S Larkey, Margaret E Connell and Nasreen Abduljaleel. Hindi CLIR in Thirty Days. *In ACM Transactions on Asian Language Information Processing*, 2003, 2(2), pp. 130-142.

[4] jinxi J and Ralph Weischedel. Cross-Lingual Retrieval for Hindi. *In ACM Transactions on Asian Language Information Processing*, Vol 2, No.1, March 2003, pp. 164-168.

[5] D Hull and G Grefenstette. Querying across languages: A dictionary-based approach to multilingual informaion retrieval. *In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, Zurich, Switzerland, pp. 49-57.

[6] Victor Lavrenko, Martin Choquette and W. Bruce Croft. Cross-Lingual Relevance Models. *In SIGIR'02*, August 11-15, 2002.

[7] Douglas W. Oard. Alternative Approaches for Cross-Language Text Retrieval. *In AAAI Symposium on Cross-Language Text and Speech Retrieval, American Association for Artificial Intelligence*, March 1997.

[8] Jinxi Xu, Ralph Weischedel and Chanh Nguyen. Evaluating a Probablistic Model for Cross-lingual Information Retrieval. *In SIGIR'01*, September 9-12, 2001.

[9] Prasad Pingali, Jagadeesh Jagarlamudi and Vasudeva Varma. Webkhoj: Indian language IR from Multiple Character Encodings. *In Internatioanl World Wide Web Conference*, May 23-26, 2006.

[10] Lisa Ballesteros and W. Bruce Croft. Resolving Ambiguity for Cross-language Retrieval. *In SIGIR'98*, Melbourne, Australia.

[11] Avinash Chopde. ITRANS version 5.30. http://www.aczone.com/itrans, July, 2001.

[12] James Allan, Victor Lavrenko and Margaret E. Connell. A Month to Topic Detection and Tracking in Hindi. *In ACM Transactions on Asian Language Processing*, Vol. 2, No.2, June 2003, pp. 85-100.

[13] Paul Clough and Mark Sanderson. Measuring Pseudo Relevance Feedback & CLIR. In SIGIR'04, July 25-29, 2004, UK.