Bengali, Hindi and Telugu to English Ad-hoc Bilingual task at CLEF 2007

Sivaji Bandyopadhyay, Tapabrata Mondal, Sudip Kumar Naskar,
Asif Ekbal, Rejwanul Haque, Srinivasa Rao Godavarthy
Department of Computer Science and Engineering,
Jadavpur University, Kolkata-700032, INDIA
sbandyopadhyay@cse.jdvu.ac.in, sivaji_cse_ju@yahoo.com

Abstract

This paper presents the experiments carried out at Jadavpur University as part of participation in the CLEF 2007 ad-hoc bilingual task. This is our first participation in the CLEF evaluation task and we have considered Bengali, Hindi and Telugu as query languages for the retrieval from English document collection. We have discussed our Bengali, Hindi and Telugu to English CLIR system as part of the ad-hoc bilingual task, English IR system for the ad-hoc monolingual task and the associated experiments at CLEF. Query construction was manual for Telugu-English ad-hoc bilingual task, while it was automatic for all other tasks.

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.2.3 [**Database Management**]: Languages-*Query Languages*

General Terms

Languages, Performance, Experimentation

Keywords

Ad-hoc cross language information retrieval, Indian languages, Bengali, Hindi, Telugu

1 Introduction

Cross-language information retrieval (CLIR) research involves the study of systems that accept queries (or information needs) in one language and return objects of a different language. These objects could be text documents, passages, images, audio or video documents. Cross-language information retrieval focused on the cross-language issues from information retrieval (IR) perspective rather than machine translation perspective.

Some of the key technical issues [7] for cross language information retrieval can be thought of as:

(i). How can a query term in one language L_1 be expressed in another language L_2?

(ii). What mechanisms determine which of the possible translations of text from L_1 to L_2 should be retained?

(iii). In cases where more than one translation is retained, how can different translation alternatives be weighed?

Many different techniques were experimented in various CLIR systems in the past in order to address these issues. These techniques can be broadly classified [3] as controlled vocabulary based and free text based systems at very high level. However, it is very difficult to create, maintain and scale a controlled vocabulary for CLIR systems in a general domain for a large corpus. Researchers came up with models that can be built on the full text of the corpus. The free text based system research can be broadly classified on the corpus-based and knowledge-based aspects. Corpus-based systems may use parallel or comparable corpora, which are aligned at word level, sentence level or passage level to learn models automatically. Knowledge-based systems might use bilingual dictionaries or ontologies, which form the handcrafted knowledge readily available for the systems to use. Hybrid systems were also built combining the knowledge-based and corpus-based approaches. Apart from these approaches, the extension of monolingual IR techniques such as vector-based models, relevance modeling techniques [14] etc., to cross language IR were also explored.

In this work we have discussed our experiments on CLIR for Indian languages to English, where the queries are in Indian languages and the documents to be retrieved are in English. Experiments were carried out using queries in three Indian languages using the CLEF 2007 experimental setup. The three languages chosen were Bengali, Hindi and Telugu, which are predominantly spoken in the eastern India, northern India and the southern India respectively.

2 Related Work

Very little work has been done in the past in the areas of IR and CLIR involving Indian languages. In the year 2003 a surprise language exercise [4] was conducted at ACM TALIP[1]. The task was to build CLIR systems for English to Hindi and Cebuano, where the queries were in English and the documents were in Hindi and Cebuano. Five teams participated in this evaluation task at ACM TALIP providing some insights into the issues involved in processing Indian language content. A few other information access systems were built apart from this task such as cross language Hindi headline generation [2], English to Hindi question answering system [13] etc. International Institute of Information Technology (IIIT), Hyderabad, built a monolingual web search engine for various Indian languages, which is capable of retrieving information from multiple character encodings [10]. In CLEF 2006 ad-hoc document retrieval task, Hindi and Telugu to English Cross Lingual Information Retrieval task [11] were reported by IIIT, Hyderabad.

Some research was previously carried out in the areas of machine translation involving Indian languages [1], [13] etc. Most of the Indian language MT efforts involve studies on translating various Indian languages amongst themselves or translating English to Indian language content. Hence most of the Indian language resources available for the works are largely biased to these tasks. Recently, Government of India has initiated a consortia project titled “Development of Cross–Lingual Information Access System”, where the query would be in any of the six different Indian languages (Bengali, Hindi, Marathi, Telugu, Tamil, Punjabi) and the output would be also in the user’s own language.

3 Our Approach

The experiments carried out by us for CLEF 2007 are based on stemming, zonal indexing and TFIDF based ranking model with bilingual dictionary look up. There were no readily available bilingual dictionaries that could be used as databases for this work, so we had to develop bilingual dictionaries from the available resources in the Internet. The method of zonal indexing was applied on the English document collection after removing stop words and performing stemming operation. The keywords in the English document collection were indexed using the n-gram indexing methodology. The query terms were extracted from the topic files using bilingual dictionaries. The Information Retrieval system was working on a TFIDF based ranking model. Query construction was carried out manually for the Telugu-English bilingual task due to the unavailability of the machine-readable Telugu-English bilingual dictionary.

3.1 Zonal Indexing and Query Construction

In zonal indexing, a particular document is divided into n number of zones/regions, say, w_1, w_2 , …… ,w_n. Then a weight is associated with each zone in such a way that the sum of all weights results in 1. Here, we divided each document into two zones, say, w_1 and w_2. The zone ‘w_1’ contains the contents of ED, PT, DK, EI, KH, HD and AU tags and ‘w_2’ region contains the contents of ID and TE tags of the Los Angeles Times (LA TIMES, 2002) documents. The weights heuristically assigned to w1 and w2 were 0.3 and 0.7 respectively. The contents of these two zones for all the documents were checked for stop words and then stemmed. Relative term frequency of a content word in a document is then calculated in each of the w_1 and w_2 regions as the ratio of the number of occurrences of the content word in the region to the total number of content words present in that region. The relative term frequencies of any content word in the two regions are normalized and added together to get the relative term frequency of that content word in the entire document. These content words which could be multiwords were used as index keywords.

We have created a list of stop-words for each language, i.e., English, Bengali, Hindi and Telugu. We have also prepared a list of words/terms that identifies whether the index terms in the narration part provided with each topic talk about relevance/irrelevance of the index terms with respect to the topic. This list has been prepared for the languages studying the

[1] ACM Transactions on Asian Language Information Processing, http://www.acm.org/pubs/talip

corresponding topic files. Stop words are first eliminated from the topic files. For every n-gram identified from the topic file all possible lower order (n-1) grams starting from unigrams were considred as query words. For example, for the trigram “Australian Prime Minister” identified from the topic file, the following were included in the query as query expansion:

Monograms
Australian
Prime

Minister

Bigrams
Australian Prime

Prime Minister

Trigram
Australian Priime Minister

3.2 Query Translation

The available Bengali-English dictionary[2] was conveniently formatted for the machine-processing tasks. The Hindi-English dictionary was developed from the available English-Bengali and Bengali-Hindi machine readable dictionaries. Initially, the English-Hindi dictionary was constructed. This dictionary was then converted into a Hindi-English dictionary for further use. A Telugu – English human readable online dictionary was used for query construction from Telugu topic files. Related works on dictionary construction can be found in [8].

The popular *Porter Stemming* [12] algorithm has been used in this work in order to remove the suffixes from the terms in the English topic file. Indian languages are inflectional / agglutinative in nature and thus demand good stemming algorithms. Due to the absence of good stemmers for Indian languages, the words in the Bengali, Hindi and Telugu topic files are subjected to suffix stripping using manually prepared suffix lists in the respective languages. The terms remaining after suffix removal are looked up in the corresponding bilingual Bengali / Hindi / Telugu to English dictionary. All English words/terms found in the Bengali / Hindi / Telugu to English dictionary for a word are considered, these may be synonyms or may correspond to different senses of the source language word. Many of the terms may not be found in the bilingual dictionary, as the term is a proper name or a word from a foreign language or a valid Indian language word, which did not occur in the dictionary. Dictionary look up may fail in some cases due to the errors involved in the process of stemming and/or suffix removal. For handling dictionary look up failure csases, a transliteration from Indian languages to English was attempted assuming the word to be most likely a proper name not to be found in the bilingual dictionaries. The transliteration engine is the modified joint source-channel model [2] based on the regular expression based alignment techniques. Three different bilingual training sets namely, Bengali-English, Hindi-English and Telugu-English were developed to train the transliteration engine. The Bengali-English training set contains approximately 25,000 bilingual examples of proper names, particularly person and location names. The Hindi-English and Telugu-English bilingual training sets were developed from the Bengali-English training set. The Hindi-English and Telugu-English training sets contain 5,000 bilingual training examples. The Indian language terms are thus translated and transliterated into the English terms accordingly. These translated/transliterated terms are then added together to form the English language query terms as part of query expansion. This algorithm for query translation and transliteration addresses the first issue of representing query in one language (L_1) to another language (L_2). The query translation process considers all the alternative translations / transliterations with equal weight.

Once the translations for the words of the Bengali, Hindi and Telugu topic files were obtained, all possible n-grams (n=1 to no. of query words in the title) were extracted for the title of each topic as explained in Section 3.1. We considered consecutive words as an n-gram, if no stop-word appears in between. For the English topic file, n-grams were extracted from title, description and narration part. For Bengali, Hindi and Telegu topic files, ngrams and all possible monograms were considered for the description and narration parts of each topic.

[2] http://dsal.uchicago.edu/dictionaries/biswas-bengali

3.3 Experiments

The evaluation document set consists of 135,917 documents from Los Angeles Times of 2002. Among these, a large number of documents were containing no element. A set of 50 topics representing the information need was given for each of the languages, Bengali, Hindi, Telugu and English. A set of human relevance judgments for these topics was generated by assessors at CLEF. These relevance judgements are binary relevance judgements and are decided by a human assessor after reviewing a set of pooled documents using the relevant document pooling technique. The system evaluation framework is similar to the Cranfield style system evaluations and the measures are similar to those used in TREC[3] [6]. Three different runs were submitted related to the three Indian languages, one for each of the three languages, Bengali, Hindi and Telugu as our task in the ad-hoc bilingual track. Another run was submitted for English as a part of the ad-hoc monolingual task. Three runs were performed using the title, description and narration parts of the topic files for Bengali, Hindi and English. Only title and description parts of the topic file were considered for the bilingual Telugu-English run.

3.4 CLEF 2007 Evaluation for Bengali-English, Hindi-English, Telugu-English Bilingual Ad-hoc Task and English Monolingual Ad-hoc Task

The run statistics for the 4 runs submitted to CLEF 2007 are described in Table 1. Clearly the geometric average precision metrics and its difference from mean average precision metrics suggests the lack of robustness in our system. There were certain topics that performed very well across the language pairs as well as for English also, but there were many topics where the performance was very low. The values of the evaluation metrics of Table 1 show that our system performs the best for the monolingual English task. As part of the bilingual ad-hoc tasks, the system performs best for the Telugu followed by Hindi and Bengali. The key to these higher values of the evaluation metrics in the Telugu-English bilingual run compared to other two bilingual runs (Hindi-English and Telugu-English) may be the manual tasks that were carried out during indexing. But it is also evident that the automatic runs for Hindi-English and Bengali-English tasks achieved a performance comparable to the manual run of Telugu-English. The overall relatively low performance of the system particularly with Indian language queries is the indicative of the fact that simple techniques such as dictionary lookup with minimal lemmatization such as suffix removal may not be sufficient for the morphologically rich Indian language CLIR. Relatively low performance of Bengali/Hindi suggests the need for broader coverage of dictionary and good morphological analyzer is inevitable for Bengali/Hindi CLIR in order to achieve a reasonable performance.

Run	**MAP**	**R-Prec**	**GAP**	**B-Pref**
Bengali Title + Description + Narration	10.18%	12.48%	2.81%	12.72%
Hindi Title + Description + Narration	10.86%	13.70%	2.78%	13.43%
Telugu Title + Description	11.28%	13.92%	2.76%	12.95%
English Title + Description + Narration	12.32%	14.40%	4.63%	13.68%

Table 1: Run Statistics

The mean precision with retrieved documents graphs are shown in figures 1.(a) – (d) for the Bengali-English, Hindi-English, Telugu-English and the monolingual English tasks. The interpolated precision with standard recall graphs is shown in figures 2.(a) – (d) for the four different runs. The figures 2.(a) – (d) suggest that the effect of ranking has not been much in the system. The sloping of curve seems to be consistent all across, as opposed to a rapid sloping for the first few recall points in all the runs. A good ranking algorithm would consistently push relevant documents to the top ranks; thereby resulting in a rapid sloping of the curve for the first few recall points.

[3] Text Retrieval Conferences, http://trec.nist.gov

4 Conclusion and Future Work

Our experiments suggest that simple TFIDF based ranking algorithms may not result in effective CLIR systems for Indian language queries. Any additional information added from corpora either resulting in source language query expansion or the target language query expansion or both could help. Machine readable bilingual dictionaries with more coverage would have improved the results. An aligned bilingual parallel corpus would be an ideal resource to have in order to apply certain machine learning approaches. Application of word sense disambiguation methods on the translated query words would have a positive effect on the result. A robust stemmer is required for the highly inflective Indian languages. We would like to automate the query construction task of Telugu in future.

5 References

[1] Akshar Bharati, Rajeev Sangal, Dipti M Sharma and Amba P Kulkarni. Machine Translation Activities in India: A Survey. In the *Proceedings of Workshop on Survey on Research and Development of Machine Translation in Asian Countries*, 2002.

[2] Asif Ekbal,, Sudip Naskar and Sivaji Bandyopadhyay. A Modified Joint Source-Channel Model for Transliteration. In Proceedings of the *COLING/ACL*, 191-198, Sydney, Australia, 2006.

[3] Bonnie Dorr, David Zajic and Richard Schwartz. Cross-language Headline Generation for Hindi. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3): 270-289, 2003

[4] Douglas Orad. Alternative Approaches for Cross Language Text Retrieval. In *AAAI Symposium on Cross Language Text and Speech Rretrieval*, USA, 1997.

[5] Douglas W. Orad. The Surprise Language Exercises. *ACM Transactions on Asian Language Information Processing* (*TALIP*), 2(2): 79-84, 2003.

[6] Ellen M. Voorchees and Donna Harman. Overview of the Sixth Text Retrieval Conferences (TTREC). In Proceedings of the *Workshop of Sixth Text Retrieval Conference*. 241-273, Morristown, NJ, USA, 1996.

[7] Gregory Grefenstette and G. Grefenstette. Cross-Language Information Retrieval. *Kluwer Academic Publishers*, Norwell, MA, USA, 1998.

[8] James Mayfield and Paul McNamee. Converting On-line Bilingual Dictionaries from Human-readable form to Machine-readable form. In Proceedings of *25th Annual International ACM SIGIR Conference on Research and Development in Informational Retrieval*, 405-406, New York, NY, USA, 2002, ACM Press.

[9] Prasad Pingali, Jagadeesh jagarlamudi and Vasudeva Varma. Webkhoj: Indian Language IR from Multiple Character Encodings. In Proceedings *of the 15th International Conference on World Wide Web*, 801-809, New York, NY, USA, 2006. ACM Press.

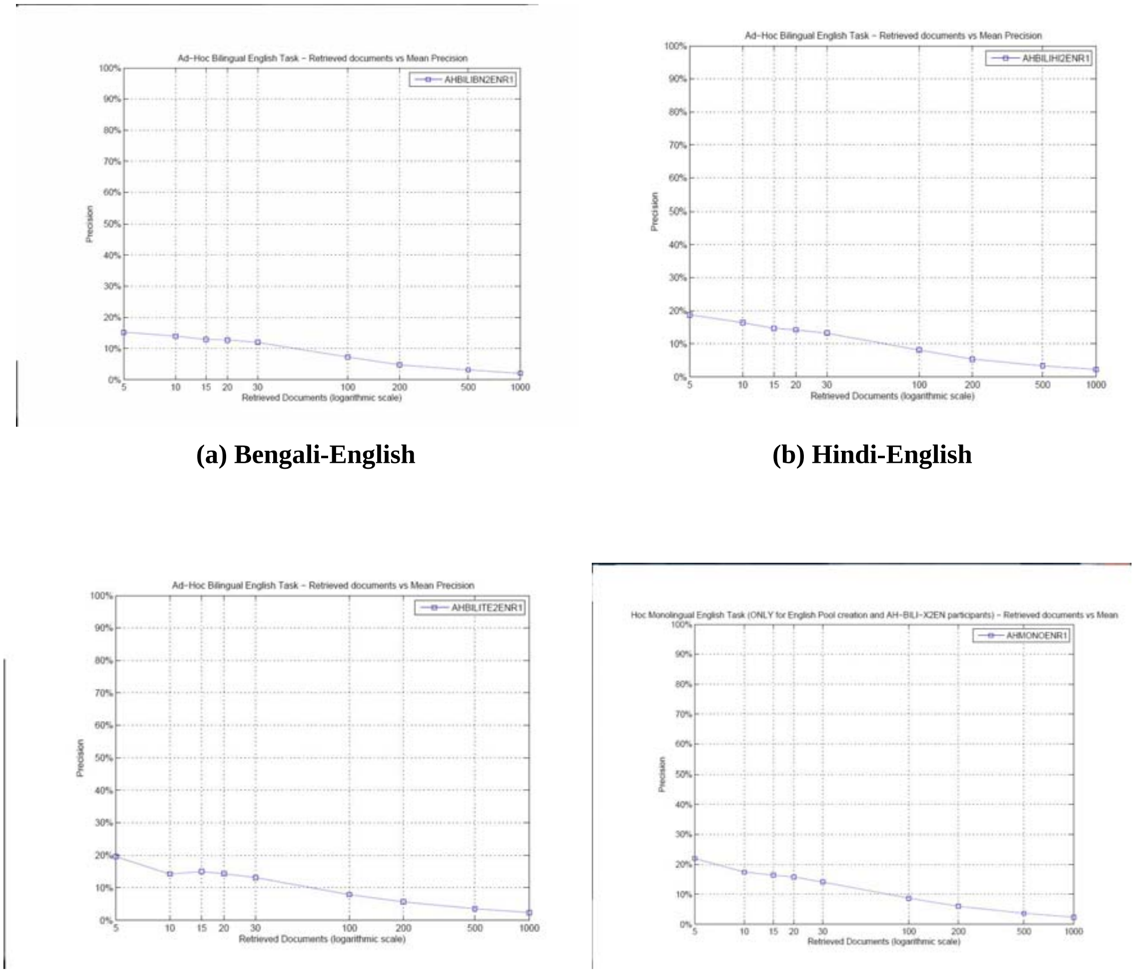
[10] Prasad Pingali, Vasudeva Varma. Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006. In *Working Notes for the CLEF 2006 Wokshop (Cross Language Adhoc Task),* 20-22 September, Alicante, Spain.

[11] Porter, M. F. (1980). An Algorithm for Suffix Stripping, *Program*, 14(3), 130-137.

[12] Satoshi Sekine and Ralph Grishman. Hindi-English Cross-Lingual Question-Answering System. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3): 181-192, 2003

[13] Sudip Naskar and Sivaji Bandyopadhyay. Use of Machine Translation in India: Current Status, In Proc. of *MT SUMMIT-X*, 465-470, Phuket, Thailand

[14] Victor Lavrenko, Martin Choquette and W. Bruce Croft. Cross-Lingual Relevance Models. In Proceedings of the *25th Annual International ACM SIGIR Conference on Research and Development in Information Retriev*al, 175-182, New York, NY, USA, ACM 2002, ACM Press.



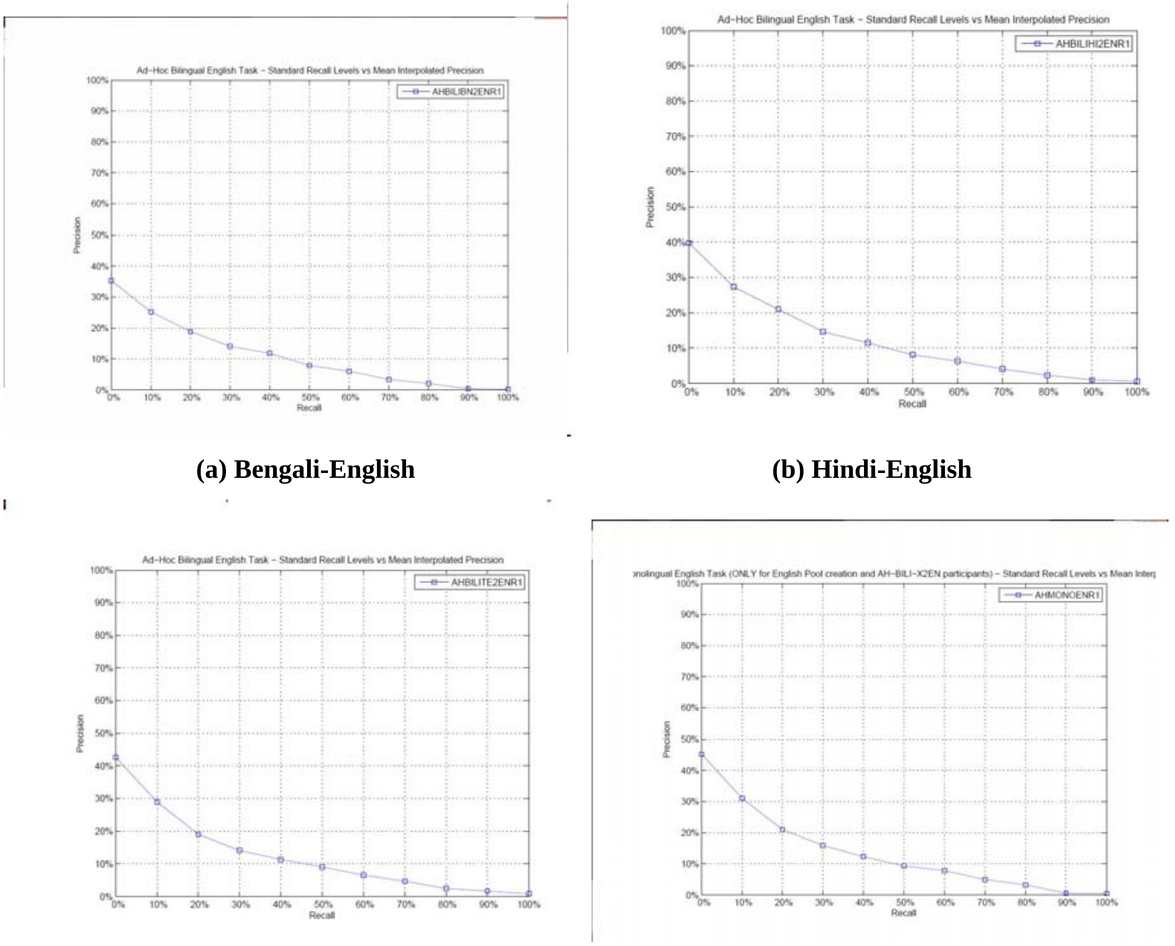
(a) Bengali-English

(b) Hindi-English

(c) Telugu-English

(d) Monolingual English

Figure 1: PRECISION VS RETRIEVED DOCUMENT (LOGARITHMIC SCALE)



(a) Bengali-English

(b) Hindi-English

(c) Telugu-English

(d) Monolingual English

Figure 2: STANDARD RECALL LEVEL VS MEAN INTERPOLATED PRECISION