Bengali and Hindi to English CLIR Evaluation

Debasis Mandal, Mayank Gupta, Sandipan Dandapat,
Pratyush Banerjee, and Sudeshna Sarkar

Department of Computer Science and Engineering
IIT Kharagpur, India - 721302
{debasis.mandal,pratyushb}@gmail.com,
{sandipan,mayank,sudeshna}@cse.iitkgp.ernet.in

Abstract. This paper presents a cross-language retrieval system for the retrieval of English documents in response to queries in Bengali and Hindi, as part of our participation in CLEF[1] 2007 Ad-hoc bilingual track. We followed the dictionary-based Machine Translation approach to generate the equivalent English query out of Indian language topics. Our main challenge was to work with a limited coverage dictionary (of coverage $\sim$ 20%) that was available for Hindi-English, and virtually non-existent dictionary for Bengali-English. So we depended mostly on a phonetic transliteration system to overcome this. The CLEF results point to the need for a rich bilingual lexicon, a translation disambiguator, Named Entity Recognizer and a better transliterator for CLIR involving Indian languages. The best MAP values for Bengali and Hindi CLIR for our experiment were 7.26% and 4.77%, which are 20% and 13% of our best monolingual retrieval, respectively.

1 Introduction

The growing number of multilingual web-accessible documents has benefitted many users who are able to read documents in more than one language. India, being a multilingual country of 22 official languages, has most of its inhabitants bilingual in nature, and exposed to English or Hindi (or both), in addition to their mother tongue. This necessitates the cross-language retrieval across the web, where the information need is expressed in a user's native language (the source language) and a ranked list of documents is returned in another language (target language). Since the language of query and documents to be retrieved are different, either the documents or queries need to be translated for CLIR. But the translation step tends to cause a reduction in the retrieval performance of CLIR as compared to monolingual information retrieval. Due to this reason, unambiguous translation is an important part of CLIR research. Various approaches involving parallel corpora, machine translation and bilingual dictionary have been experimented to address this problem [1,2,3,4]. However, in this paper we will restrict ourselves in the dictionary-based Machine Translation approach.

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

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 95–102, 2008.

Bengali and Hindi are considered to be very resource-poor languages, in the sense that few language resources or tools (e.g. bilingual lexicon, morphological generator, parser etc) are available for them. This is due to the reason that much work has not yet been done in CLIR involving them. The other obstacle is the percentage of web contents for these languages, which is much less compared to other resource-rich languages, like English. Even within this limited content we faced several language-specific obstacles, like proprietary encodings of much of the web texts, that prohibited us to build the required training corpus for these languages. The scarcity of good parallel corpora restricted us to build the computational resources, like bilingual statistical lexicon and statistical transliterator. Moreover, the stemmers that were available for these languages usually make use of an extensive set of linguistic rules and thus lack comprehensive coverage. Furthermore, a named entity recognizer for Bengali and Hindi were also not available during the experiments [5].

Under this limited resource scenario, the sole objective of our participation in CLEF was to evaluate the basic CLIR system we had for Bengali and Hindi[2], and to explore the resource dependency, sources of improvement and comparability with other CLIR systems. This was our first participation in CLEF and we conducted six bilingual and three monolingual retrieval experiments for two language pairs: Hindi and Bengali to English. The investigation of CLEF evaluation results provided the necessary scope of improvement in the system and the importance of various IR components in great detail.

The rest of the paper is organized as follows. Section 2 presents some of the primitive and influencing works in CLIR involving Indian languages. The following section builds our CLIR model on the basis of bilingual lexicons, stemmers and transliterators. CLEF evaluations of our experiments and their analysis 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 involving Indian languages is relatively a new area of research among Natural Language Processing community, and first major work involving Hindi occurred only during TIDES Surprise Language exercise [7] in 2003. The objective of the exercise was to retrieve Hindi documents, provided by LDC (Linguistic Data Consortium), in response to queries in English. Interestingly, it was just an evolving field in India at that time and so no Indian university participated in the exercise. The five participants displayed a beautiful collaboration among them and submitted individual systems within one month period. They built a statistical lexicon out of parallel corpora [6,8,9] and used it to design Machine Translation based cross-lingual systems. The experiments had many interesting outcomes. Assigning TF.IDF weights on query terms and expanding query using training corpora were shown to improve the cross-language results even over

[2] Hindi and Bengali are world's fifth and seventh most spoken languages, respectively. Ethnologue: Languages of the World, 15^{th} ed. (2005) `http://www.ethnologue.com`

Hindi monolingual runs [8]. Larkey et al. approached the problem using Language modeling approach [6] and showed the importance of a good stemmer for highly inflected languages, like Hindi. Finally, the exercise established the need of a good bilingual lexicon, query normalization, stop-words removal, stemming, query expansion with feedback and transliteration for the good result for Hindi.

The recent interest in cross-language research has given rise to a consortium for Cross-Language Information Access (CLIA) involving six Indian languages and premier research institutes across the country. As part of the ongoing research, several approaches have been tested and evaluated for CLIR in Indian languages in CLEF. The Language modeling coupled with Probabilistic transliteration, used by Larkey et al. [6] in surprise exercise, was also shown to be fruitful for Hindi and Telugu to English CLIR by Prasad et al. [10]. The approach also showed a significant improvement in performance over the simple dictionary-based Machine Translation. Manoj et al. performed Marathi and Hindi to English CLIR using Iterative Disambiguation Algorithm, which involves disambiguating multiple translations based on term-term co-occurrence statistics [11]. Jagadeesh et al. [12] had used a word alignment table, learned by a Statistical Machine Translation (SMT) system and trained on aligned parallel sentences, to convert the query into English. Sivaji et al. [13] has approached the problem for Hindi, Bengali and Telugu languages using a zonal-indexing approach on the corpus documents. In their approach, each document was first divided into some zones and then assigned some weights, the relative frequency of a term is then calculated based on zonal frequencies and thereafter used as an index keyword for query generation. Some of the other issues with the CLIA involving Indian languages and their feasible remedies are also discussed in [14,15,16].

3 Experiments

A basic dictionary-based Machine Translation approach, viz., tokenization, stop-words removal, stemming, bilingual dictionary look up and phonetic transliteration were followed to generate the equivalent English query out of Indian language topics. The main challenge of our experiment was to transliterate out-of-dictionary words properly and use limited bilingual lexicon efficiently. We had access to a Hindi-English bilingual lexicon[3] of $\sim$ 26K Hindi words, a Bengali biochemical lexicon of $\sim$ 9K Bengali words, a Bengali morphological analyzer and a Hindi Stemmer. In order to achieve a successful retrieval under this limited resource, we adopted the following strategies: Structured Query Translations, phoneme-based followed by a list-based named entity transliterations, and performing no relevance judgment. Finally, the English query was fed into Lucene search engine and the documents were retrieved along with their normalized scores, which follows the Vector Space Model (VSM) of Information Retrieval. Lucene was also used for the tokenization and indexing of corpus documents.

[3] 'Shabdanjali'.
http://ltrc.iiit.net/onlineServices/Dictionaries/Dict_Frame.html

3.1 Structured Query Translation

After stemming of the topic words, the stemmed terms were looked up in the machine readable bilingual lexicon. If the term occurred in the dictionary, all of the corresponding translations were used to generate the final query. Parts-of-speech information of the topic words were not considered during translation. But many of those terms did not occur in the lexicon due to following reasons: limitations of the dictionary, improper stemming, the term is a foreign word or a named entity [10]. A close analysis showed that only 13.47% of the terms from 'title+desc' fields and 19.59% of the terms from 'title+desc+narr' fields were only found in the Hindi bilingual lexicon. For Bengali bilingual lexicon, the probability of finding a term dropped to below 5%.

3.2 Query Transliteration

The out-of-dictionary topic words were then transliterated into English using a phonetic transliteration system, assuming them to be *proper nouns*. The system works in the character level and converts every single Hindi or Bengali character in order to transliterate a word. But it produced multiple possibilities for every word, since English is not a phonetic language. For example, the Hindi term for *Australia* had four possible transliterations as output: *astreliya*, *astrelia*, *austreliya*, and *austrelia*. To disambiguate the transliterations, the terms were then matched against a manually-built named entity list with the help of an approximate string matching algorithm, *edit-distance algorithm*. The algorithm returns the best match of a term for pentagram statistics. For above example, the list correctly returned *Australia* as the final query term in cross-language runs.

Note that we did not expand the query using Pseudo Relevance Feedback (PRF) system. This is due to the fact that it sometimes does not improve the overall retrieval significantly for CLIR, rather hurts the performance by increasing noise towards the end retrievals [17]. Furthermore, it also increases the number of queries for which no relevant documents are returned, as shown in [8].

4 Results

The objective of Ad-Hoc Bilingual (X2EN) English task was to retrieve at least 1000 documents corresponding to each of the 50 queries from English target collection and submit them in ranked order. The data set and metrics for the Ad-Hoc track evaluation are described in detail in [18]. To evaluate the performance of our cross-language retrieval system, six bilingual runs were submitted for Bengali and Hindi, as shown in Table 1[4]. As a baseline, we also submitted three monolingual English runs consisting of various topic fields. For each of the Indian languages, the comparisons are made with respect to the best base run, viz., monolingual 'title+desc' run. The best values of Recall and MAP (Mean Average Precision) for the base run are 78.95% and 36.49%, respectively.

[4] The DOI corresponding to a <Run ID> is http://dx.doi.org/10.2415/AH-BILI-X2EN-CLEF2007.KHARAGPUR.<Run ID>

Table 1. Cross-language runs submitted in CLEF 2007

Sl.#	Run ID	Topic Lang	Topic Field(s)
1	BENGALITITLE	Bengali	title
2	BENGALITITLEDESC	Bengali	title+desc
3	BENGALITITLEDESCNARR	Bengali	title+desc+narr
4	HINDITITLE	Hindi	title
5	HINDITITLEDESC	Hindi	title+desc
6	HINDITITLEDESCNARR	Hindi	title+desc+narr

The results of our cross-language task are summarized in Table 2 and Table 3. Table 2 shows that the recall gradually improved with the addition of more relevant terms from the topic fields for Bengali, as expected, but the same did not repeat for Hindi. This result was a surprise to us as we had used a bilingual lexicon of superior performance for Hindi. A careful analysis revealed that the value of MAP is also poorer for Hindi, as seen in Table 3, contrary to our expectation. Moreover, variations in the values of MAP and R-precision over different topic fields are not much for Hindi, as compared to Bengali. However, the precision values with respect to top 5, 10 and 20 retrievals demonstrate a steady increase for each of them.

Table 2. Summary of bilingual runs of the Experiment

Run ID	Relevant Docs	Relevant Retrieved	Recall (in %)	% mono	B-Pref
BENGALITITLE	2247	608	27.60	34.96	5.43
BENGALITITLEDESC	2247	851	37.87	47.97	10.38
BENGALITITLEDESCNARR	2247	906	40.32	51.07	11.21
HINDITITLE	2247	708	31.51	39.91	9.95
HINDITITLEDESC	2247	687	30.57	38.72	11.58
HINDITITLEDESCNARR	2247	696	30.97	39.23	12.02

The anomalous behavior of Hindi can be explained in terms of translation disambiguation during query generation. Query wise score breakup revealed that the queries with more named entities always provided better results than their counterparts. With the increase of lexical entries and Structured Query Translation (SQT), more and more 'noisy words' were incorporated into final query in the absence of any translation disambiguation algorithm, thus bringing down the overall performance. The average English translations per Hindi word in the lexicon were 1.29, with 14.89% Hindi words having two or more translations. For example, the Hindi word 'rokanA' (to stop) had 20 translations, making it highly susceptible towards noise. Figure 1 shows the frequency distribution of dictionary entries with their corresponding number of translations in the Hindi

bilingual dictionary. It is also evident from Table 3 that adding extra information to query through 'desc' field increases the performance of the system, but adding 'narr' field has not improved the result significantly. The post-CLEF analysis revealed that this field constituted two parts: relevance and irrelevance, and was meant to prune out the irrelevant documents during retrieval. But we did not make any effort in preventing the irrelevant retrieval in our IR model.

Table 3. Precision results (in %) for bilingual runs in CLEF 2007

Run ID	MAP	% mono	R-Prec	P@5	P@10	P@20
BENGALITITLE	4.98	13.65	5.86	4.80	6.60	7.00
BENGALITITLEDESC	7.26	20.00	8.53	10.00	10.20	8.80
BENGALITITLEDESCNARR	7.19	19.70	9.00	11.60	10.80	10.70
HINDITITLE	4.77	13.07	5.34	8.40	6.40	5.40
HINDITITLEDESC	4.39	12.03	5.19	9.20	8.60	7.10
HINDITITLEDESCNARR	4.77	13.07	5.76	10.40	8.40	7.30

The results in Table 3 show that the best MAP values for Bengali and Hindi CLIR for our experiment are 7.26% and 4.77% which are 20% and 13% of our best base run, respectively. Although the result of Bengali is comparable (10.18%) with only other participant for the language in CLEF 2007 [13], the results for Hindi in our experiment was much poorer than the best entry (29.52%) [11]. Lack of a good bilingual lexicon can be attributed as the primary reason for our poor result.

The other shortcoming of our system is the homogeneous distribution of precision with respect to retrieved documents and interpolated recall, as evident from Figure 2. This clearly demands for a good feedback system (e.g. Pseudo Relevance Feedback) to push the most relevant documents to the top. Apart from the costly query refinement operation, improvement can also be made by

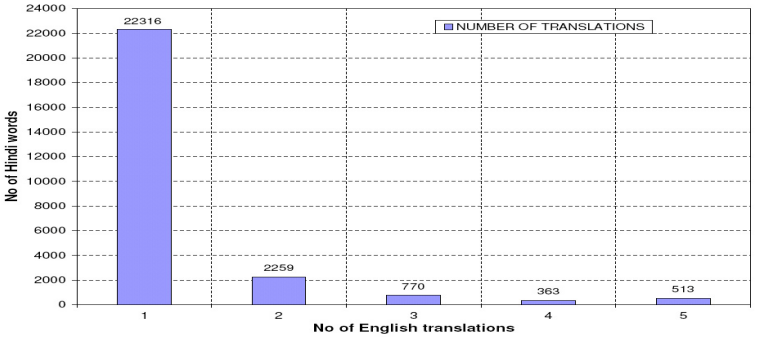

Fig. 1. Frequency distribution of number of translations in Hindi bilingual dictionary

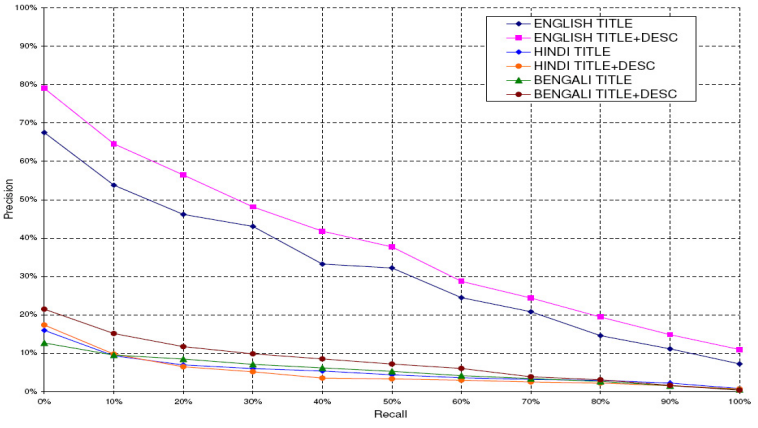

Fig. 2. Recall vs. Precision results for the Experiment

identifying the named entities in the query and assigning them a higher relative weight with respect to other query terms.

5 Conclusions and Future Works

This paper described an experiment of Bengali and Hindi to English cross-language text retrieval as part of CLEF 2007, its evaluation results and few post-evaluation analyses. The poorer performance of our system with respect to other resource-rich participants clearly pointed out the necessity of a rich bilingual lexicon, a good transliteration system, and a relevance feedback system. Further, part of speech (POS) information will help to disambiguate the translations. Performance of the stemmer also has an important role in cross-language retrieval for morphologically rich languages, like Bengali and Hindi.

Our future work includes building named entity recognizers and efficient transliteration system based on statistical and linguistic rules. We would also like to analyze the effect of feedback system in cross-language query expansion. Language modeling is another approach we would like to test upon for a better cross-language retrieval involving Indian languages.

Acknowledgment

We would like to thank Mr. Sunandan Chakraborty of the Department of Computer Science & Engineering, IIT Kharagpur, for his generous help to resolve various programming issues during integration of the system.

References

1. Hull, D., Grefenstette, G.: Querying across languages: A dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19^{th} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 49–57 (1996)
2. Diekema, A.R.: Translation Events in Cross-Language Information Retrieval. ACM SIGIR Forum 38(1) (2004)
3. Bertoldi, N., Federico, M.: Statistical Models for Monolingual and Bilingual Information Retrieval. Information Retrieval 7, 53–72 (2004)
4. Monz, C., Dorr, B.: Iterative Translation Disambiguation for Cross-Language Information Retrieval. In: SIGIR 2005, Salvador, Brazil, pp. 520–527 (2005)
5. Mandal, D., Dandapat, S., Gupta, M., Banerjee, P., Sarkar, S.: Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
6. Larkey, L.S., Connell, M.E., Abduljaleel, N.: Hindi CLIR in Thirty Days. ACM Transactions on Asian Language Information Processing (TALIP) 2(2), 130–142 (2003)
7. Oard, D.W.: The surprise language exercises. ACM Transactions on Asian Language Information Processing (TALIP) 2(2), 79–84 (2003)
8. Xu, J., Weischedel, R.: Cross-Lingual Retrieval for Hindi. ACM Transactions on Asian Language Information Processing (TALIP) 2(1), 164–168 (2003)
9. Allan, J., Lavrenko, V., Connell, M.E.: A Month to Topic Detection and Tracking in Hindi. ACM Transactions on Asian Language Processing (TALIP) 2(2), 85–100 (2003)
10. Pingali, P., Tune, K.K., Varma, V.: Hindi, Telugu, Oromo, English CLIR Evaluation. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730. Springer, Heidelberg (2007)
11. Chinnakotla, M.K., Ranadive, S., Bhattacharyya, P., Damani, O.P.: Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
12. Jagarlamudi, J., Kumaran, A.: Cross-Lingual Information Retrieval System for Indian Languages. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
13. Bandyopadhyay, S., Mondal, T., Naskar, S.K., Ekbal, A., Haque, R., Godavarthy, S.R.: Bengali, Hindi and Telugu to English Ad-hoc Bilingual task at CLEF 2007. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
14. Pingali, P., Jagarlamudi, J., Varma, V.: Webkhoj: Indian language IR from Multiple Character Encodings. In: International World Wide Web Conference (2006)
15. Pingali, P., Varma, V.: IIIT Hyderabad at CLEF 2007 Adhoc Indian Language CLIR task. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)
16. Pingali, P., Varma, V.: Multilingual Indexing Support for CLIR using Language Modeling. In: Bulletin of the IEEE Computer Society Technical Committee on Data Engineering (2007)
17. Clough, P., Sanderson, M.: Measuring Pseudo Relevance Feedback & CLIR. In: SIGIR 2004, UK (2004)
18. Nunzio, G.M.D., Ferro, N., Mandl, T., Peters, C.: CLEF 2007: Ad-Hoc Track Overview. In: Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2007 Workshop (2007)