Cross-Lingual Information Retrieval System for Indian Languages

Jagadeesh Jagarlamudi and A. Kumaran

Multilingual Systems Research,
Microsoft Research India,
Bangalore, India
{jags,a.kumaran}@microsoft.com
http://research.microsoft.com/india

Abstract. This paper describes our attempt to build a Cross-Lingual Information Retrieval (CLIR) system as a part of the Indian language sub-task of the main Adhoc monolingual and bilingual track in CLEF[1] competition. In this track, the task required retrieval of relevant documents from an English corpus in response to a query expressed in different Indian languages including Hindi, Tamil, Telugu, Bengali and Marathi. Groups participating in this track were required to submit a English to English monolingual run and a Hindi to English bilingual run with optional runs in rest of the languages. Our submission consisted of a monolingual English run and a Hindi to English cross-lingual run.

We used a word alignment table that was learnt by a Statistical Machine Translation (SMT) system trained on aligned parallel sentences, to map a query in the source language into an equivalent query in the language of the document collection. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. On the CLEF 2007 data set, our official cross-lingual performance was 54.4% of the monolingual performance and in the post submission experiments we found that it can be significantly improved up to 76.3%.

1 Introduction

The rapidly changing demographics of the internet population [1] and the plethora of multilingual content on the web [2] has attracted the attention of the Information Retrieval(IR) community to develop methodologies for information access across languages. Since the past decade [3,4,5,6], researchers are looking at ways to retrieve documents in a language in response to a query in another language. This fundamentally assumes that users can read and understand documents written in foreign language but are unable to express their information need in that language. There are arguments against this assumption as well: for example, Moulinier and Schilder argue that it is unlikely that the information in another language will be useful unless users are already fluent in that language [7]. However, we argue that in specific cases such methodologies could still be

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

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 80–87, 2008.

valid. For example, in India students learn more than one language from their childhood and more than 30% of the population can read and understand Hindi apart from their native language [8]. The multilingual capability of the users exhibits a great demand for systems with the capability to retrieve relevant documents in languages different from the language in which information need is expressed.

Lack of resources is still a major reason for relatively less number of efforts in the cross-lingual setting in the Indian subcontinent. Research communities working in Indian Languages, especially on Machine Translation (MT) [9], have built some necessary resources like morphological analyzer and bilingual dictionaries for some languages. As we demonstrate in this paper, even though these resources are built mainly for MT, they can still be used as a good starting point to build a CLIR system. More specifically, in this paper we will describe our first attempt at building a CLIR system using a bilingual statistical dictionary that was learnt automatically during the training phase of a SMT [10] system.

In the rest of the paper we will first define the problem in section 2, followed by a brief description of our approach in section 3. In section 4 we will describe data set along with the resources used and also present the results reported in the CLEF competition (sec. 4.1). This section also includes some analysis of the results. Section 5 presents conclusion and outlines our plans for future work.

2 Problem Statement

We have participated in the Indian Language sub-task of the main CLEF 2007 Ad-hoc monolingual and bilingual track. This track tests the performance of systems in retrieving relevant documents in response to a query in the same and different language from that of the document set. In the Indian language track, documents are provided in English and queries are specified in different languages including Hindi, Telugu, Bengali, Marathi and Tamil. The system has to retrieve 1000 relevant documents as a response to a query in any of the above mentioned languages. All the systems participating in this track are required to submit an English to English monolingual run and a Hindi to English bilingual run. Runs in the rest of the languages are optional. We have submitted an English to English monolingual and Hindi to English bilingual run.

3 Approach

Converting the information expressed in different languages to a common representation is inherent in cross-lingual applications to bridge the language barrier. In CLIR, either the query or the document or both need to be mapped onto the common representation to retrieve relevant documents. Translating all documents into the query language is less desirable due to the enormous resource requirements. Usually the query is translated into the language of the target collection of documents. Typically three types of resources are being exploited for translating the queries: bilingual machine readable dictionaries, parallel texts

and machine translation systems. MT systems typically produce one candidate translation thus some potential information which could be of use to IR system is lost. Even though researchers [11] have also explored considering more than one possible translation to avoid the loss of such useful information, another difficulty in using the MT system comes from the fact that most of the search queries are very short and thus lack necessary syntactic information required for translation. Hence most approaches use bilingual dictionaries.

In our work, we have used statistically aligned Hindi to English word alignments that were learnt during the training phase of a machine translation system. The query in Hindi is translated into English using word by word translation. For a given Hindi word, all English words which have translation probability above a certain threshold are selected as candidate translations. Only top 'n' of these candidates are selected as final translations in order to reduce ambiguity in the translation. This process may not produce any translations for some of the query words because either the word is not available in the parallel corpus or all translation probabilities are less than the threshold. In such cases, we attempt to transliterate the query word into English. We have used a noisy channel model based transliteration algorithm [12]. The phonemic alignments between Hindi characters and corresponding English characters are learnt automatically from a training corpus of parallel names in Hindi and English. These alignments along with their probabilities are used, during viterbi decoding, to transliterate a new Hindi word into English. As reported, this system will output the correct (fuzzy match) English word in top 10 results, with an accuracy of about 30%(80%). As a post processing step, target language vocabulary along with approximate string matching algorithms like soundex and edit distance measure [13] were used to filter out the correct word from the incorrect ones among the possible candidate transliterations.

Once the query is translated into the language of the document collection, standard IR algorithms can be used to retrieve relevant documents. We have used Langauge Modeling [14] in our experiments. In a Language Modeling framework, both query formulation and retrieval of relevant documents are treated as simple probability mechanisms. Essentially, each document is assumed to be a language sample and the query to be a sample from the document. The likelihood of generating a query from a document ($p(q|d)$) is associated with the relevance of the document to the query. A document which is more likely to generate the user query is considered to be more relevant. Since a document, considered as a bag of words is very small, compared to the whole vocabulary, most of the times the resulting document models are very sparse. Hence smoothing of the document distributions is very crucial. Many techniques have been explored and because of its simplicity and effectiveness we chose the relative frequency of a term in the entire collection to smooth the document distributions.

In a nutshell, structural query translation [6] is used to translate query into English. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. The following section describes our approach applied in the CLEF 2007 participation and some further experiments to calibrate the quality of our system.

4 Experiments

In both the Adhoc bilingual 'X' to English track and Indian language sub track, the target document collection consisted of 135,153 English news articles published in Los Angeles Times, from the year 2002. During indexing of this document collection, only the text portion (embedded in <LD> and <TE> tags) was considered. Note that the results reported in this paper do not make use of other potentially useful information present in the document, such as, the document heading (with in <DH> tag) and the photo caption (in <CP> tag), even though we believe that including such information would improve the performance of the system. The resulting 85,994 non-empty documents were then processed to remove stop words and the remaining words were reduced into their base form using Porter stemmer [15].

The query set consists of 50 topics originally created in English and translated later into other languages. For processing Hindi queries, a list of stop words was formed based on the frequency of a word in the monolingual corpus corresponding to the Hindi part of the parallel data. This list was then used to remove any less informative words occurring in the topic statements. The processed query was then translated into English using a word alignment table.

We have used a word alignment table that was learnt by the SMT [10] system trained on 100K Hindi to English parallel sentences acquired from Webdunia to translate Hindi queries. Since these alignments were learnt for machine translation purpose, the alignments included words in their inflectional forms. For this reason we have not converted the query words into their base form during the translation. Table 1 shows the statistics about the coverage of the alignment table corresponding to different levels of threshold on the translation probability (column 1), note that a threshold value of 0 corresponds to having no threshold at all. Columns 2 and 3 indicate the coverage of the dictionary in terms of source and target language words. The last column denotes the average number of English translations for a Hindi word. It is very clear and intuitive that as the threshold increases the coverage of the dictionary decreases. It is also worth noting that as the threshold increases the average translations per source word decreases, indicating that the target language words which are related to the source word but not synonymous are getting filtered.

4.1 Results

For each query, a pool of candidate relevant documents is created by combining the documents submitted by all systems. From this pool assessors filter out actual

Table 1. Coverage statistics of the word alignment table

Threshold	Hindi words	English words	Translations per word
0	57555	59696	8.53
0.1	45154	54945	4.39
0.3	14161	17216	1.59

Table 2. Monolingual and Cross-lingual experiments

	Monolingual		Crosslingual	
	LM(td)	LM(tdn)	LM(td)	LM(tdn)
MAP	0.3916	0.3964	0.1994	0.2156
p@10	0.456	0.454	0.216	0.294

relevant documents from the non relevant ones. These relevance judgements are then used to automatically evaluate the quality of participating CLIR systems.

Here we present the official results of our monolingual English run and Hindi to English bilingual run. In our case, we specifically participated in only one Indian language - Hindi, though the data was available in 5 Indian languages. For our official submission, with the aim of reducing noise in the translated query, we used a relatively high threshold of 0.3 for the translation probability. To avoid ambiguity, when there are many possible English translations for a given Hindi word, we included only the two best possible translations according to the translation table. Table 2 shows the official results of our submission. We have submitted different runs using title, description (td) and title, description and narration (tdn) as query.

In a second set of experiments, we experimented with the effect of various levels of threshold and the number of translations above the threshold on Mean Average Precision(MAP) score. The results obtained by the cross-lingual system with varying threshold are compared against the monolingual system in fig. 1. The right most bar in each group represents the monolingual performance of the system. The figure shows that in each group the performance increases as the threshold decrease and it decreases if you consider more number of

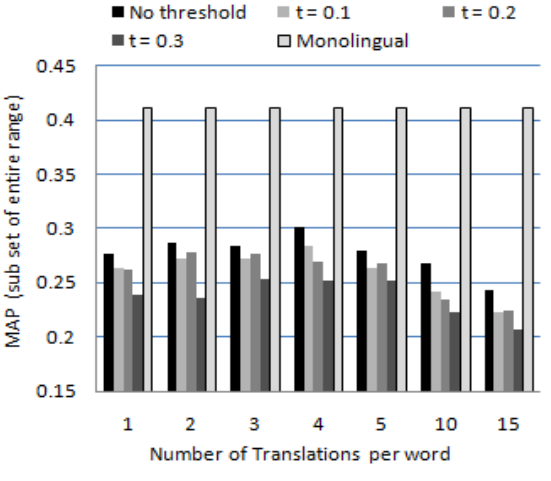

Fig. 1. Hindi to English cross-lingual performance with varying levels of threshold on translation probability

Table 3. Fraction of translated words

Threshold	Translated words
0	0.803
0.1	0.7892
0.2	0.711
0.3	0.6344

Table 4. Skewedness of the dictionary

No. of top words (% of vocab size)	% of dictionary entries
10 (0.0173)	3.44
20 (0.0347)	5.87
50 (0.0869)	11.05
100 (0.173)	17.01

possible translations per word (drop when 10 and 15 translations were considered) perhaps due to the shift in query focus with the inclusion of many less synonymous target language words. For the CLEF data set, we found that considering the four most possible translations without any threshold (left most bar) of the translation probability gave us the best results (73.4% of monolingual IR performance).

As the threshold decreases potentially two things can happen: words which were not translated previously can get translated or new target language words whose translation probability was below the threshold earlier will now become part of the translated query. Table 3 shows the fraction of query words that were translated corresponding to each of these thresholds. Table 3 and figure 1 show that as the threshold on the translation probability decreases, the fraction of query words getting translated increases, resulting in an overall increase in system performance. But the performance increase between having no threshold and a threshold of 0.1 compared to the small fraction of new words that got translated suggest that even noisy translations, even though they are not truly synonymous, might help CLIR. This is perhaps due to the fact that for the purposes of IR, in absense of appropriate translation having a list of associated words may be sufficient to identify the context of the query [16].

During the query wise analysis we found that cross-lingual retrieval performed marginally better than the monolingual system on five topics, whereas it under performed on 19 queries. The analysis of 11 queries, on which the CLIR performance is worse than 10% of monolingual, revealed three kinds of errors as very prominent: words missing in the translation dictionary, inappropriate selection of target language word and transliteration errors. Even though the dictionary coverage looks very exhaustive, queries still have relatively more number of content words which are not covered by the dictionary. This is mainly because the dictionary since it is learned statistically, it is skewed. Table 4 shows the number of top words, in terms of having multiple target language translations, along with the fraction of dictionary entries corresponding to them. The top 100 words

corresponding to a small fraction of 0.173% of vocabulary cover almost 17% of the dictionary and most of these words are less informative for Information Retrieval purpose. This is evident as the dictionary is learnt statistically, where a frequent word has more chances of being aligned to many target language words depending on the context in which it is being used.

If a query contains a relatively more frequent word, since it has more chances of being aligned to many words, it may become a source of noise in the translated query. This has also been the cause for the selection of inappropriate target language words, even when you consider only top 'n' translations. In order to select the appropriate target language word that is more common in the target language corpus, we performed a simple adaptation of the translation dictionary. Instead of using $p(e|h)$ directly we have replaced it with $p(h|e) \cdot p(e)$ and normalized appropriately to follow the probability constraints. The unigram probability counts for the English words are computed from the target language corpus instead of the English part of parallel sentences. We assumed that this will prefer words that are more frequent in the collection to words that are relatively infrequent. Such a simple technique has resulted in an improvement of 4.23% indicating a scope for further exploration.

5 Conclusion and Future Work

This paper describes our first attempt at building a CLIR system with the help of a word alignment table learned statistically. We present our submission in the Indian language sub-task of the Adhoc monolingual and bilingual track of CLEF 2007. In post submission experiments we found that, on CLEF data set, a Hindi to English cross-lingual information retrieval system using a simple word by word translation of the query with the help of a word alignment table was able to achieve $\sim$ 76% of the performance of the monolingual system. Empirically we found that considering four most probable word translations with no threshold on the translation probability gave the best results.

In our analysis, we found the coverage of dictionary and the choice of appropriate translation to be the potential places for improvement. In the future we would like to exploit either the parallel or comparable corpora for selecting the appropriate translation of a given source word. We would also like to compare the distribution statistics of a statistically learned dictionary with respect to a hand crafted dictionary of similar size to compare them for CLIR purposes.

References

1. Internet, `http://www.internetworldstats.com`
2. GlobalReach, `http://www.global-reach.biz/globstats/evol.html`
3. Ballesteros, L., Croft, W.B.: Dictionary methods for cross-lingual information retrieval. In: Thoma, H., Wagner, R.R. (eds.) DEXA 1996. LNCS, vol. 1134, pp. 791–801. Springer, Heidelberg (1996)

4. Hull, D.A., Grefenstette, G.: Querying across languages: A dictionary-based approach to Multilingual Information Retrieval. In: SIGIR 1996: Proc. of the 19th annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 49–57. ACM Press, New York (1996)
5. McNamee, P., Mayfield, J.: Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In: SIGIR 2002: Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 159–166. ACM Press, New York (2002)
6. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. Information Retrieval 4(3-4), 209–230 (2001)
7. Moulinier, I., Schilder, F.: What is the future of multi-lingual information access?. In: SIGIR 2006 Workshop on Multilingual Information Access 2006, Seattle, Washington, USA (2006)
8. Burkhart, G.E., Goodman, S.E., Mehta, A., Press, L.: The Internet in India: Better times ahead?. Commun. ACM 41(11), 21–26 (1998)
9. Bharati, A., Sangal, R., Sharma, D.M., Kulakarni, A.P.: Machine Translation activities in India: A survey. In: Workshop on survey on Research and Development of Machine Translation in Asian Countries (2002)
10. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. 29(1), 19–51 (2003)
11. Kwok, K.L., Choi, S., Dinstl, N.: Rich results from poor resources: Ntcir-4 monolingual and cross-lingual retrieval of korean texts using chinese and english. ACM Transactions on Asian Language Information Processing (TALIP) 4(2), 136–162 (2005)
12. Kumaran, A., Kellner, T.: A generic framework for machine transliteration. In: SIGIR 2007: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 721–722. ACM Press, New York (2007)
13. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: English Translation in Soviet Physics Doklady, pp. 707–710 (1966)
14. Ponte, J.M., Croft, W.B.: A Language Modeling Approach to Information Retrieval. In: ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281 (1998)
15. Porter, M.F.: An algorithm for suffix stripping. Program: News of Computers in British University libraries 14, 130–137 (1980)
16. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Inf. Process. Manage. 43(4), 866–886 (2007)