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

Abstract

This paper describes our first participation in the Indian language sub-task of the main Adhoc monolingual and bilingual track in CLEF[1] competition. In this track, the task is to retrieve 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 are required to submit a English to English monolingual run and a Hindi to English bilingual run with optional runs in rest of the languages. We had submitted 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 source language into an equivalent query in the language of the target document collection. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. On 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 73.4%.

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 Managment**]: Languages—*Query Languages*

General Terms

Experiments, Performance

Keywords

Information Retrieval, Cross-lingual Information Retrieval, Statistical Machine Translation, CLEF, Bilingual Dictionary, Query Translation

1 Introduction

The rapidly changing demographics of the internet population [7] and the plethora of multilingual content on the web [5] has attracted the attention of Information Retrieval(IR) community to

[1] http://www.clef-campaign.org

develop methodologies for cross-lingual information accessing. Since the past decade [1, 6, 11, 14] 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 unable to express their information need in that language. There are arguments against this assumption as well: For example, [12] argues that it is unlikely that the information in another language will be useful unless users are already fluent in that language. However, we argue that in specific cases such methodologies could still be 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 [4]. This situation exhibits great utility 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 Indian subcontinent. Research communities working in Indian Languages, especially on Machine Translation (MT) [2], have built some necessary resources like morphological analyzer and bilingual dictionaries for some languages. Even though these resources are built mainly for MT, they can still be used as a good starting point to build a Cross-Lingual Information Retrieval (CLIR) system, as we demonstrate in this paper. More specifically, in this paper we will describe our first attempt in building a CLIR system using the bilingual statistical dictionary that was learnt automatically during the training phase of a SMT [13] system.

The rest of the paper is organized as follows. We will first define the problem in section 2, followed by the presentation of our approach in section 3. In sections 4.1 & 4.2 we will describe data set along with the resources used and present the performance of our system (section 4.3) in the CLEF competition. Section 4 includes some analysis of the results. Section 5 presents our conclusion and identifies our plans for future work.

2 Problem Statement

Cross Language Evaluation Forum (CLEF) aims at promoting research in the design of multi-lingual, multi-modal retrieval systems by providing an opportunity for the research communities working in different languages to collaborate and share their experiences. Each year it organizes a series of evaluation tracks to test different aspects of cross-language information retrieval system development.

We have participated in the CLEF competition, specifically 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 the 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 response to a query in any of the above mentioned languages. All the systems participating in this track are required to submit a English to English monolingual run and a Hindi to English bilingual run. Runs in rest of the languages are optional.

3 Approach

Converting the information expressed in different languages to a common representation is inherent to cross-lingual applications to build the language barrier. In CLIR, either the query or the document or both need to be mapped into the common representation to retrieve the 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 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. Researchers [9] have also explored considering more than one possible translation to avoid the loss of useful information. Another difficulty in using the MT system comes from the fact that most of the search queries are very short and lack necessary syntactic information required for translation. Hence most approaches use bilingual dictionaries.

In our work, we have used a statistically aligned Hindi to English word alignments that were learnt during the training phase of machine translation. The query in Hindi language is translated into English using word by word translation. For a given Hindi word, all English words which have translation probability above certain threshold are selected as candidate translations. Only top 'n' of these candidates are selected as final translations to reduce ambiguity in the translation. The aligned bilingual dictionary may not contain some of the query words because either the word is not available in the parallel corpus or the 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 [8]. 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%). Target language vocabulary along with approximate string matching algorithms like soundex and edit distance measure [10] 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 the relevant documents. We have used Langauge Modeling [15] in our experiments. In Language Modeling framework, both query formulation and retrieval of relevant documents are treated as simple probability mechanisms. Essentially, each document is assumed as a language sample and query as a sample from this document. The likelihood of generating a query from the 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 bag of words is very small when compared to 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 [14] is used to translate query into English. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. Following section describes our approach applied in the CLEF 2007 participation and some further experiments to calibrate the quality of our system.

4 Experiments

4.1 CLEF Data set for Adhoc track

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 the indexing of this document collection, only text portion (embedded in <LD> and <TE> tags) was considered. Note that the results reported in this paper does 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 an information would improve the performance of the system. The resulting 85,994 non-empty documents were then processed to remove the stop words and the remaining words were reduced into their base form using Porter stemmer [16].

50 topics originally created in English and translated later into other languages were distributed among the participants. For processing Hindi queries, a list of stop words was formed based on the frequency of word in the monolingual corpus obtained 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 the word alignment table.

4.2 Word Alignment Table as Bilingual Dictionary

We have used the word alignment table that was learnt by the SMT [13] system trained on 100K Hindi to English parallel sentences to translate Hindi queries. Since these alignments were learnt primarily for machine translation purpose, the alignments included words that occurs 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 correspond 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 denote 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.

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 1: Coverage statistics of the word alignment table

4.3 Results

Each of the participating systems was required to submit 1000 relevant documents for each topic. For each query, a pool of candidate relevant documents is created by combining the documents submitted by all systems. From the pool of such candidate documents assessors filter out actual relevant documents from non relevant ones. These relevance judgements are then used to automatically evaluate the quality of cross-lingual retrieval of participating CLIR systems.

In this section we discuss the 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 have 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 allowed only 2 best possible translations according to the statistical alignments learnt by SMT. 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. Narration seems to be improving the cross-lingual retrieval performance, in terms of Mean Average Precision (MAP), more than that in monolingual setting. In the rest of the experiments it is assumed that narration is included as a part of the query unless explicitly mentioned. Figure 1 shows a comparison of cross-lingual and mono-lingual submissions in terms of precision at different levels of interpolated recall.

	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

Table 2: monolingual and cross-lingual experiments.

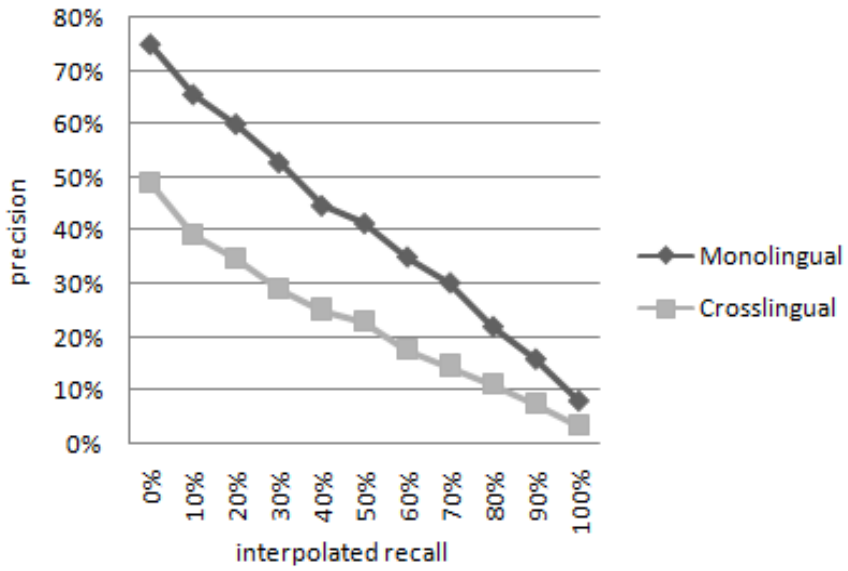

Figure 1: Precision at different levels of interpolated recall.

In a second set of experiments, we experimented with various levels of threshold and the number of translations above the threshold and their effect on MAP score. The results obtained by monolingual system and cross-lingual system with varying threshold are shown plotted in Figure 2. The x-axis represents the number of top words considered when many of the target language words have translation probabilities above the chosen threshold. Note that y-axis represents only a subset (0.15-0.45) of the entire possible range (0-1) in which a MAP score can lie. The right most bar represents the monolingual performance of the system. The figure shows that the performance increases as the threshold decrease but again it drops if you consider all the possible translations (dip 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 4 most possible translations with out any threshold (left most bar) on the translation probability gave us the best results (73.4% of monolingual IR performance).

Threshold	Translated words
0	0.803
0.1	0.7892
0.2	0.711
0.3	0.6344

Table 3: Fraction of translated words

As the threshold decrease 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 will now become the candidates 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 2 show that as the threshold on the translation probability decrease, 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 true synonymous, might help CLIR. This perhaps due to the fact that for the purposes of IR, having a list of associated words may be sufficient to identify the context of the query [3].

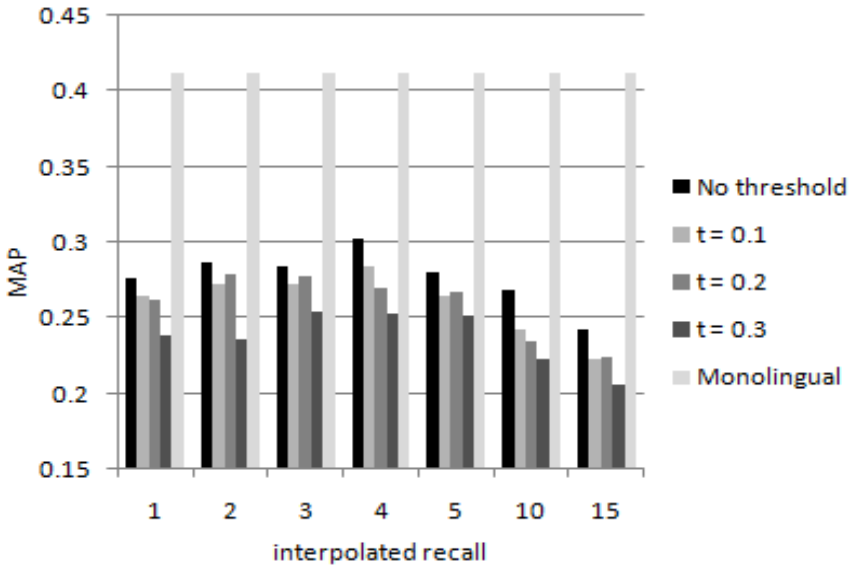

Figure 2: Comparison of the CLIR Performance with varying threshold on translation probability.

5 Conclusion and Future Work

This paper describes our first attempts in building a CLIR system with the help of a word alignment table learned, from a parallel corpora, primarily for statistical machine translation. We present our experience and results of our participation 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 $\sim$ 73% of the performance of the monolingual system. Empirically we found that considering 4 most probable word translations with no threshold on the translation probability gave the best results.

Since the quality of the dictionary will affect the performance of the system, in future we would like to explore the effect of size and quality of the parallel data on the word alignments, and subsequently on the CLIR performance. We would also like to compare the use of a statistically learned word alignments with respect to a hand crafted dictionary of similar size for CLIR application.

References

[1] Lisa Ballesteros and W. Bruce Croft. Dictionary methods for cross-lingual information retrieval. In *DEXA '96: Proceedings of the 7th International Conference on Database and Expert Systems Applications*, pages 791–801, London, UK, 1996. Springer-Verlag.

[2] Akshar Bharati, Rajeev Sangal, Dipti M Sharma, and Amba P Kulakarni. Machine Translation activities in India: A survey. In *Workshop on survey on Research and Development of Machine Translation in Asian Countries*, 2002.

[3] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Inf. Process. Manage.*, 43(4):866–886, 2007.

[4] Grey E. Burkhart, Seymour E. Goodman, Arun Mehta, and Larry Press. The internet in india: better times ahead? *Commun. ACM*, 41(11):21–26, 1998.

[5] GlobalReach. *http://www.global-reach.biz/globstats/evol.html.*

[6] David A. Hull and Gregory Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57, New York, NY, USA, 1996. ACM Press.

[7] Internet. *http://www.internetworldstats.com.*

[8] A. Kumaran and Tobias Kellner. A generic framework for machine transliteration. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 721–722, New York, NY, USA, 2007. ACM Press.

[9] Kui Lam Kwok, Sora Choi, and Norbert Dinstl. 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.

[10] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *English translation in Soviet Physics Doklady*, pages 707–710, 1966.

[11] Paul McNamee and James Mayfield. Comparing cross-language query expansion techniques by degrading translation resources. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 159–166, New York, NY, USA, 2002. ACM Press.

[12] Isabelle Moulinier and Frank Schilder. What is the future of multi-lingual information access? In *SIGIR 2006 Workshop on Multilingual Information Access 2006*, Seattle, Washington, USA, 2006.

[13] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, 2003.

[14] Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Järvelin. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Inf. Retr.*, 4(3-4):209–230, 2001.

[15] Jay M. Ponte and W. Bruce Croft. A Language Modeling Approach to Information Retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

[16] M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.