Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006

Prasad Pingali and Vasudeva Varma
Language Technologies Research Centre
IIIT, Hyderabad, India
pvvpr@iiit.ac.in, vv@iiit.ac.in

Abstract

This paper presents the experiments of Language Technologies Research Centre (LTRC)[1] as part of their participation in CLEF[2] 2006 ad-hoc document retrieval task. This is our first participation in the CLEF evaluation tasks and we focused on Afaan Oromo, Hindi and Telugu as query languages for retrieval from English document collection. In this paper we discuss our Hindi and Telugu to English CLIR system and the experiments at CLEF.

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

Measurement, Performance, Experimentation

Keywords

Ad-hoc cross language text retrieval, Indian languages, 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 the machine translation (MT) perspective. The motivation for a separate research into such systems was that CLIR was not merely coupling of IR and MT, and a lot of processing usually performed in machine translation systems may not be necessary for CLIR. Also on the other hand, machine translation systems rely on syntactically well formed sentences as input to the system, which may not be a realistic assumption for an IR system, as most of the IR queries tend to be very short and many times without any syntactic correctness and hence very little context to perform syntactic parsing or disambiguate automatically. However, some times keyword based queries might also contain valid phrases which could be the level of language syntax one could rely on for CLIR systems.

[1]LTRC is a research centre at IIIT, Hyderabad, India. http://ltrc.iiit.ac.in
[2]Cross Language Evaluation Forum. http://clef-campaign.org.

Some of the key technical issues [3] for cross language information retrieval can be thought of as

- How can a query term in L_1 be expressed in L_2?
- What mechanisms determine which of the possible translations of text from L_1 to L_2 should be retained?
- In cases where more than one translation are retained, how can different translation alternatives be weighed?

In order to address these issues, many different techniques were tried in various CLIR systems in the past. These techniques can be broadly classified [7] as controlled vocabulary based and free text based systems at a 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. Therefore very quickly researchers realized it would be essential to come up with models that can be built from the full text of the corpus. The free text based system research can be broadly classified on the corpus-based and knowledge-based aspects. This classification comes from the type of information resources used by the CLIR systems in order to address the above mentioned issues. For example, knowledge based systems might use bilingual dictionaries or ontologies which form the hand-crafted knowledge readily available for the systems to use. On the other hand corpus-based systems may use parallel or comparable corpora which are aligned at word level, sentence level or passage level to learn models automatically. 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 [5] etc., to cross language IR were also explored.

In this paper we discuss 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 conducted using queries in two Indian languages using the CLEF 2006 experimental setup. The two languages chosen were Hindi which is predominantly spoken in north India and Telugu which is predominantly spoken in southern part of India. In the rest of the paper we discuss CLIR and related work in these Indian languages and also our own experiments at CLEF 2006.

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 [8] was conducted at ACM TALIP[3]. 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 [11] etc. We previously built a monolingual web search engine for various Indian languages which is capable of retrieving information from multiple character encodings [10]. However, no work was found related to CLIR involving Telugu or any other Indian language other than Hindi.

Some research was previously done in the areas of machine translation involving Indian languages [1]. Most of the Indian language MT efforts involve studies on translating various Indian languages amongst themselves or translating English into Indian language content. Hence most of the Indian language resources available for our work are largely biased to these tasks. This led to the challenge of using resources which enabled translation from English to Indian languages for a task involving translation from Indian languages to English.

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

3 Problem Statement

The problem statement of CLIR task discussed in this paper is as defined in the ad-hoc track of CLEF 2006. The ad-hoc track tests mono- and cross-language textual document retrieval. The bilingual task on target collections in English would test systems where the topics are supplied in a variety of languages including Amharic, Afaan Oromo, Hindi, Telugu and Indonesian. In this paper we discuss our system for Hindi and Telugu languages therefore the system will be provided with a set of 50 topics in Hindi and Telugu where each topic represents an information need for which English text documents need to be retrieved and ranked. An example topic in Telugu would look as shown below.

<top>
<num> C302 </num>
<TE-title> వినియోగదారుల బహిష్కరణ </TE-title>
<TE-desc>
వినియోగదారుల బహిష్కరణలను చూపే, వివరించే పత్రాలను చూపుము
</TE-desc>
<TE-narr>
సంబంధిత పత్రాలు వినియోగదారుల బహిష్కరణలు, వినియోగదారుల బహిష్కరణల్లో కూడి ఉన్న నీతి నియమాలు కూడా చర్చిస్తాయి. వినియోగదారుల బహిష్కరణలే సంబంధితము, రాజకీయ బహిష్కరణలను వదిలెయ్యాలి.
</TE-narr>
</top>

Each topic comes with a unique number identifying the topic, a title, a description and a narrative. A title is typically a few words in length and is characteristic of a real world IR query. The description of a topic contains more detailed description of what the user is looking for, as a natural language statement. A narrative contains a little more information than the description in the sense that it also give additional information of what is relevant and what is not relevant. Such information would be very useful for systems which use both relevance as well as irrelevance information into their models. The system should use these topics as input or manually a set of keywords can be generated by a human and provided to the system. In this paper we restrict our problem to automatically retrieving the relevant documents with the input topics. The system is expected to provide an output of 1000 documents for each topic in a ranked order which are evaluated against a set of manually created relevance judgements. The possible judgements for each retrieved documents could either be relevant or irrelevant. In other words the relevance judgements are binary.

4 Our Approach

Our submission to CLEF 2006 uses a vector based ranking model with bilingual lexicon using word translations combined with a set of heuristics for query refinement after translation. The ranking is achieved using a vector based ranking model using TFIDF ranking algorithm. We used the lucene framework to index the English documents. All the English documents were stemmed and stop words were eliminated to obtain the index terms. These terms were indexed using the Lucene[4] search engine using the TFIDF similarity metric.

4.1 Query Translation

The only resources we had access to were English-Hindi and English-Telugu cross language dictionaries[5] which were primarily used in English to Indian language machine translation research. The English-Hindi dictionary was conviniently formatted for machine processing, however the English-Telugu dictionary was a digitized version of a human readable dictionary. In order to convert the human readable dictionary to machine processable form, a set of regular expressions were used.

[4] http://lucene.apache.org
[5] http://ltrc.iiit.net/onlineServices/Dictionaries/

Similar approaches were previously tried to convert human readable dictionaries into a form easily processable by machines [6]. We removed a set of standard high frequency suffixes both from the queries and dictionaries before-hand. The set of prefixes we used for Hindi are similar to those mentioned in [4]. For Telugu, we used the set of suffixes as shown in Table 1.

వాడు, డు, ం, ము, ఆలు, లు, ఆల, ఆలలో, లలో, లను, లకు, ల్లో, లో, ఆనికి, కి, కు, మొక్క, చేత, తో, చే, ఆ, ఇ, ఈ, ఉ, ఊ, ఎ, ఏ, ఒ, ఓ, ఔ, ఈయ

Table 1: Telugu suffixes (full vowels may be replaced with short vowel equivalents)

The terms remaining after suffix removal are looked up in bilingual dictionary which is a English to Indian language dictionary. A set of multiple English meanings for a given query term would be obtained for a given Indian language term. Many of the terms may not be found in the bilingual lexicon since the term is a proper name or a word from a foreign language or a valid Indian language word which just did not occur in the dictionary. In some of the cases dictionary lookup for a term might also fail because of improper stemming or suffix removal. Indian languages are agglutinative languages, especially Telugu is highly agglutinative which would demand a good stemming algorithm. However, due to lack of availability of such a resource we used suffix removal technique with a set of high frequency suffixes. For lookup failure cases where the word was a proper name, a transliteration from Indian language to English was attempted. The transliteration was first performed with a set of phoneme mappings between Indian langauge and English. While this technique might succeed in a few cases, in many cases this may not transliterate into the right English term. Therefore we used approximate string matching algorithms of the obtained transliteration against the lexicon from the corpus. We used the double metaphone [9] algorithm as well as Levenstein's approximate string matching algorithm to obtain possible transliterations for the query term which was not found in the dictionary. The intersection set from these two algorithms were added to the translated English query terms. This algorithm for query translation and transliteration addresses the first issue of representing query in language L_1 in language L_2 as was previously mentioned in section 1.

4.2 Query Refinement and Retrieval

Once the translation and transliteration tasks are performed on the input Hindi and Telugu queries, we tried to address the second issue for CLIR from our list as mentioned in section 1. We tried to prune out the possible translations for the query in an effort to reduce the possible noise in translations. In order to achieve this, we used a pseudo-relevance feedback based on the top ranking documents above a threshold using the TFIDF retrieval engine. The translated English query was issued to the lucene search engine and a set of top 'n' documents were retrieved. The translated English terms that did not occur in these documents were pruned out in an effort to reduce the noisy translations. We chose 'n' to be 10 documents to refine the translated query. The final query after refinement process was issued to the lucene search engine to obtain the top 1000 TFIDF ranked documents. As evident from our approach, no efforts were made to identify the irrelevant documents in the search process. For this reason we did not use the narrative information in the topics for any of our runs. It is also evident that we did not make any efforts to weigh the various terms in the possible translations which is the third issue for CLIR as mentioned in section 1.

5 Experiments and Discussion

The evaluation document set consists of 113,005 documents from Los Angeles Times of 1994 and 56,472 documents from Glasgow Herald of 1995. A set of 50 topics representing the information need were given in Hindi and Telugu. A set of human relevance judgements for these topics were generated by assessors at CLEF. These relevance judgements are binary relevance judgements and

Run	MAP	R-Prec	GAP	B-Pref
Hindi Title	12.32%	13.14%	2.40%	12.78%
Hindi Title + Description	12.52%	13.16%	2.41%	10.91%
Telugu Title	8.09%	8.39%	0.34%	8.36%
Telugu Title + Description	8.16%	8.42%	0.36%	7.84%

Table 2: Run Statistics

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[6] [12]. Four runs were submitted related to the Indian languages, two with Hindi queries and two with Telugu queries. A run was performed using only title and a run was performed using both title and description for each of these languages.

5.1 CLEF 2006 Evaluation for Hindi-English and Telugu-English CLIR

The run statistics for the 4 runs submitted to CLEF 2006 are described in table 2. Clearly the geometric average precision metrics and its difference from mean average precision metrics here suggests the lack of robustness in our system. There were certain topics that performed very well while there were many topics where the performance was very low. Interestingly not much difference exists between the title runs and the title-description runs. This is surprising and suggests that most of the information for CLIR is coming from the titles. Another interesting finding with respect to the title and title-description runs is the ranking of these metrics. Clearly b-pref measure (which is a more recent metric) suggests that the title run was superior to title-description run, however all the other three metrics, the mean average precision, R-precision and geometric average precision suggest that the title-description run was better than the title run. The overall relatively low performance of the system with Indian language queries when compared to Afaan Oromo (an Ethiopian language we experimented with), suggests that a number of topics have low performance statistics, which was also evident from the topicwise score breakups. This is indicative of the fact that simple techniques such as dictionary lookup with minimal lemmatization such as suffix removal may not be sufficient for Indian language CLIR. Also we can observe that Telugu has performed much lower than Hindi suggesting the need for a good stemming and stem dictionaries. It was found from a previous corpora analysis of Indian languages that the number of unique words found in equal sized corpus of Hindi and Telugu, suggests that Telugu language is more agglutinative than Hindi, resulting in 4 times more number of unique words than Hindi. This suggests that the need for broader coverage of dictionary and good morphological analyzer is inevitable for Telugu CLIR in order to achieve a reasonable performance.

The interpolated recall with average precision graphs as shown in figure 1 suggests that the effect of ranking has not been much in the system. The sloping of the curve seems to be consistent all across, as opposed to a rapid sloping for the first few recall points. 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. Relatively a slight effect of ranking can be found for Hindi when compared to Telugu, however the overall indication seems to be that the ranking function is not an effective ranking mechanism.

6 Conclusion and Future Work

Our experiments suggest that simple extensions of vector based algorithms such as TFIDF 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 target language query

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

expansion or both could help. An aligned bilingual parallel corpus would be an ideal resource to have in order to apply certain machine learning approaches. However monolingual corpora based learning can be still tried as a backoff. Also need for better Indian language processing is clearly evident from our experiments. We plan to extend our system using monolingual Indian language corpora to enable source language query expansion and also factor in target language query expansion.

References

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

[2] 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.

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

[4] Leah S. Larkey, Margaret E. Connell, and Nasreen Abduljaleel. Hindi CLIR in thirty days. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):130–142, 2003.

[5] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. Cross-lingual relevance models. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 175–182, New York, NY, USA, 2002. ACM Press.

[6] James Mayfield and Paul McNamee. Converting on-line bilingual dictionaries from human-readable to machine-readable form. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 405–406, New York, NY, USA, 2002. ACM Press.

[7] Douglas Oard. Alternative approaches for cross language text retrieval. In *AAAI Symposium on Cross Language Text and Speeck Retrieval*, USA, 1997.

[8] Douglas W. Oard. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):79–84, 2003.

[9] L. Philips. The Double-Metaphone Search Algorithm. *C/C++ User's Journal*, 18(6), 2000.

[10] Prasad Pingali, Jagadeesh Jagarlamudi, and Vasudeva Varma. Webkhoj: Indian language ir from multiple character encodings. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 801–809, New York, NY, USA, 2006. ACM Press.

[11] 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.

[12] Ellen M. Voorhees and Donna Harman. The text retrieval conferences (trecs). In *Proceedings of a workshop on held at Baltimore, Maryland*, pages 241–273, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

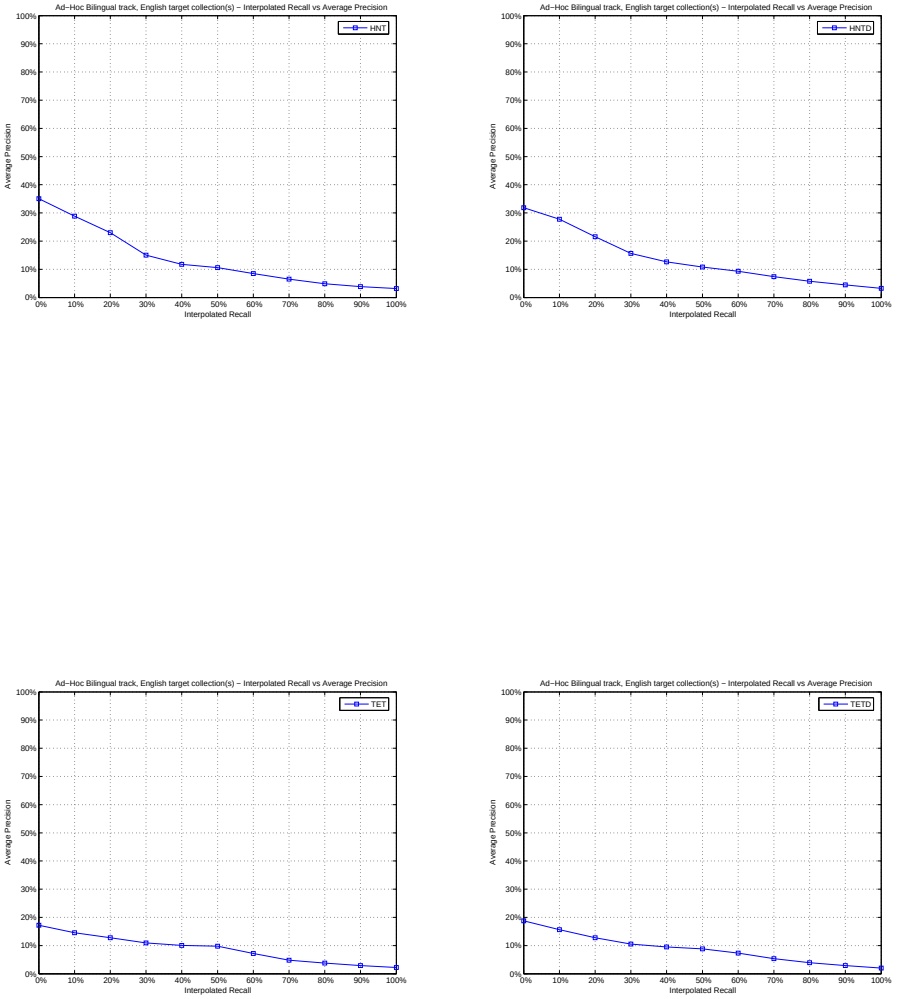

Figure 1: Interpolated Recall vs Average Precision for Hindi-English and Telugu-English runs. HNT run using Hindi title, HNTD run using Hindi title and description, TET run using Telugu title, TETD run using Telugu title and description.