IIIT Hyderabad at CLEF 2007 - Adhoc Indian Language CLIR task

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] 2007 Indian language to English ad-hoc cross language document retrieval task. In this paper we discuss our Hindi and Telugu to English CLIR system and the experiments using CLEF 2007 dataset. We used a variant of TFIDF algorithm in combination with a bilingual lexicon for query translation. We also explored the role of a document summary in fielded queries and two different boolean formulations of query translations. We find that a hybrid boolean formulation using a combination of boolean AND and boolean OR operators improves ranking of documents. We also find that simple disjunctive combination of translated query keywords results in maximum recall.

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

---

[1]LTRC is a research centre at IIIT, Hyderabad, India. http://ltrc.iiit.ac.in

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

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.

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 [5] 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 [4] 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 2007 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 2007.

2 Related Work

Very little work has been done in the past in the areas of IR and CLIR involving Indian languages. We participated in CLEF 2006 Indian language to English CLIR task, which is to the best of our knowledge, the first evaluation of CLIR where queries were provided in Hindi and Telugu and the documents to be retrieved were in English. In the year 2003 a surprise language exercise [6] 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 [10] etc. We previously built a monolingual web search engine for various Indian languages which is capable of retrieving information from multiple character encodings [7]. However, no work was found related to CLIR involving Telugu or any other Indian language other than Hindi.

---

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

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 Problem Statement

The problem statement of CLIR task discussed in this paper is as defined in the ad-hoc track of CLEF 2007. 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, Bengali, Marathi, Tamil 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.

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 2007 uses a vector based ranking model with bilingual lexicon using word translations. Out of vocabulary words or OOVs are handled using a probabilistic algorithm as mentioned in [8]. Document retrieval is achieved using an extended boolean model where queries are constructed using boolean operators among keywords and the occurrence of keywords in various types of metadata is given a different weight. The various fields that were used while retrieving the documents are mentioned in section 4.2. The ranking is achieved using a vector based ranking model using a variant of 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

A query translation need not be a true translation of the given source language query. In other words, the target language output produced need not be well formed and human readable and the only goal of such a translation is to obtain a topically similar translation in the target language to enable proper retrieval. In our system, a given source language query is translated using word

---

[4]http://lucene.apache.org

by word translation. However, the design of our system does support phrasal or multi-word expression lookup as well. Each source language word is looked up in the bilingual dictionaries for exact match as well as all words having the same prefix as the given source language word. If a given source language word does not occur in the bilingual dictionary, one character at each time is removed from the end of the word until a matching word with same prefix is found in the dictionary. This heuristic for dictionary lookup helps in translating source language words even if their morphological variants or compound words are present in the dictionary.

4.1.1 Dictionaries

From our CLEF 2006 experiments [9] we found that, using existing dictionaries which are built from human readable dictionaries may not yield good results in a CLIR task. We therefore created a new dictionary for Telugu-English and Hindi-English language pairs. While creating the dictionaries, we used the TFIDF measure of a Telugu or Hindi word from a large corpus collected from a monolingual web search engine [7]. The words having higher TFIDF would be chosen first while manually creating entries into the dictionary. The motivation for such a choice is based on our findings that dictionaries are never complete and the optimal way of building a new dictionary would be by maximizing the probability of finding the most appropriate concepts for a given cross language query word. A group of about five native language speakers of Telugu and Hindi with a reasonably good knowledge of English were chosen to create the dictionaries. The task was defined as to quickly create bilingual dictionaries for as many source language words as possible in seven days. We created a Hindi-English dictionary with 5,175 entries and Telugu-English dictionary with 26,182 entries. These entries might contain variants of a same source language word, since we do not make any effort to construct only a root word dictionary. The words chosen are as-is from the index of a web-search engine, ordered by their TFIDF scores. The guidelines given to the dictionary creators while creating the dictionaries were

- To key-in utmost five English keywords for each source language word

- To key-in entries only for the words where the creator's confidence level was very high

- To choose words in such a way that they would be topically related to the source language word and need not be an exact synonym

The third guideline mentioned above was found to help the most in building highly suitable dictionaries for a CLIR task. The idea was to assume a user giving a source language keyword for searching and imagine the most likely keywords found in a relevant target language document. In other words the created dictionary may not be an exact synonym dictionary, but a dictionary which returns topically similar keywords in the target language. Such dictionaries can also be automatically built using large comparable bilingual corpora. However, due to absence of such a resource we manually created the above said dictionaries. Also, since the word list was chosen from an unstemmed word list of a web search engine, it may also contain proper names, foreign language words or morphological variants of words as separate entries. In many such cases the ditionary creator may end up keying in the same English keywords for all the morphological variants of a given source language word.

4.2 Indexing, Retrieval and Ranking

Traditionally IR systems use the notion of fields to treat different types of metadata related to a document with different weights. Queries are constructed to search for keywords and weigh them using some prior weights assigned intuitively to a given metadata type. For example, the title of a document can be viewed as a metadata of the given article and a keyword found in the title might be deemed to be more important as against one found in the document's body. In our system, we used three such fields for each document. Two of the fields *title, body* were explicitly provided in the given corpus, while we derived a new field called *summary* of a document and chose the first 50 words of the document body as a *summary*. Different boost scores were given to each of

these fields such that *title* of the document is deemed most important followed by *summary*, then followed by the *body* of the document. In order to provide different weights to terms based on fields, a combination of term and the field name is treated as a unique entry in the inverted index. In other words, a given term occuring in both title and body are treated as two different terms and their term frequencies and document frequencies are computed individually.

As mentioned in the beginning of this section, our retrieval approach is to translate the given source language query keywords using a bilingual dictionary. We translate one term at a time and do not handle multi-word expressions and phrases in queries. However, the algorithm itself does not require any modification to be able to handle multi-term queries. It would suffice if the multi-word expressions along with their meanings are also stored in the same dictionary as the single word dictionary. Our lookup algorithm first tries to lookup entries containing longest source language expressions. This is achieved since the lookup program also internally represents the dictionary using an inverted index data structure.

Once the source language queries are translated and transliterated, the resultant English keywords used to construct boolean queries using boolean AND and OR operators. Assume the index model to contain the set of fields/metadata as $F = f_1, f_2...f_m$ and the source language query $S = s_1, s_2, ..., s_n$, and every source language keyword s_i results in multiple target language keywords. Let t_{ij} be the jth translation of source language keyword s_i. In our experiments we primarily construct a disjunctive type query Q_{disj} and a hybrid query Q_{hyb} for every k^{th} field from F as

$$Q_{disj,k} = w_k . \bigcup_{i,j} t_{ij} \tag{1}$$

$$Q_{hyb,k} = w_k . \bigcap_i \bigcup_j t_{ij} \tag{2}$$

where w_k is the boost weight given to the kth field. Finally the multiple field queries are again combined using a boolean OR operator. We report various runs based on the boolean operations on the queries and the fields on which retrieval is performed in the evaluation section.

It is evident from our approach that we do not make any 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 and treat all the translated keywords to be equally likely.

5 Experiments and Discussion

The evaluation document set consists of news articles and reports from Los Angeles Times of 2002. 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 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[5] [11]. Two runs were submitted related to the Indian languages, two with Hindi queries and two with Telugu queries. A monolingual run was also submitted to CLEF. After CLEF released the relevance judgements we conducted some more experiments. Table 5 describes the various runs we report in this section. MONO and MDISJ are the monolingual English runs. TETD, TNOSUM and TDISJ are Telugu-English CLIR runs, while HITD, HNOSUM and HDISJ are Hindi-English CLIR runs. These runs differ primarily in the fields that are used for searching and the way in which the translated keywords are combined using boolean operators. The third column in table 5 describes each of these runs.

---

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

Table 1: Run Descriptions

Run ID	Language Pair	Description
MONO	English	Monolingual run, using title, body and summary fields. Title keywords are combined using hybrid query as described in the previous section. CLEF official submission.
MDISJ	English	Monolingual run, using title, body and summary fields. All keywords are combined using boolean OR operator.
TETD	Telugu - English	Uses title, body and summary fields. Title keywords are combined using boolean AND across translations. CLEF official submission.
HITD	Hindi - English	Uses title, body and summary fields. Title keywords are combined using boolean AND across translations. CLEF official submission.
TNOSUM	Telugu - English	Title and body fields are used. Title keywords are combined using boolean AND.
HNOSUM	Hindi - English	Title and body fields are used. Title keywords are combined using boolean AND.
TDISJ	Telugu - English	Only body text is used. All translated keywords combined using boolean OR.
HDISJ	Hindi - English	Only body text is used. All translated keywords combined using boolean OR.

5.1 CLEF 2007 Evaluation for Hindi-English and Telugu-English CLIR

The average metrics for each of the runs mentioned in table 5 are described in table 5.1. The first column in table 5.1 mentiones the metric and the remaining of each column represents the various runs. Each row gives a comparison of a given metric across all the runs. The metrics listed are as provided by the TREC evaluation package [6]. Of these metrics, we find *num_rel_ret, map, bpref* and *P5* values to be interesting. Apart from these metrics we also report the number of relavent documents (num_rel), mean reciprocal rank (MRR) and interpolated recall at various precision levels (ircl_prn) metrics. From the run statistics it can be observed that the Hindi-English CLIR performs reasonably well even when the dictionary is very small around 5,000 words. Also from *P5* it can be observed that systems using boolean AND operator with appropriate boosting of metadata results in better ranking. However, such systems result in lower recall. It can be observed from *num_rel_ret* of disjunctive runs MDISJ, TDISJ and HDISJ that the system is able to retrieve more number of relavent documents when queries are combined using boolean OR operator. It can also be observed that using a summary as a metadata in retrieval might help when the translation quality is low. This fact can be observed from better performance of HNOSUM run for Hindi, which performs better than HITD. However, use of a summary results in lower performance when the translation quality is better, which can be observed from TETD and TNOSUM.

6 Conclusion and Future Work

Our experiments suggest that the performance of a CLIR system heavily depends on the type and quality of the resources being used. While the underlying IR model is also important and can play a role in the quality of a CLIR output, we found the main contribution to performance coming from the ability to convert a source language information need into the target language. We showed that, by using simple techniques to quickly create dictionaries, one can maximize

---

[6]TREC provides a trec_eval package for evaluating IR systems.

Table 2: Run Statistics

METRIC/RUN	MDISJ	MONO	TETD	TNOSUM	TDISJ	HITD	HNOSUM	HDISJ
num_q	50	50	50	50	50	50	50	50
num_ret	50000	50000	50000	50000	50000	50000	50000	50000
num_rel	2247	2247	2247	2247	2247	2247	2247	2247
num_rel_ret	1952	1629	1275	1456	1517	958	1132	1123
map	0.4003	0.3687	0.2155	0.2370	0.2170	0.1560	0.1432	0.1331
gm_ap	0.3335	0.2526	0.0834	0.0950	0.0993	0.0319	0.0321	0.0321
R-prec	0.4084	0.3979	0.2467	0.2702	0.2478	0.1689	0.1557	0.1566
bpref	0.3980	0.3850	0.2868	0.3083	0.2750	0.2104	0.2026	0.2005
recip_rank	0.7161	0.6584	0.5160	0.4814	0.5419	0.3778	0.3270	0.3416
ircl_prn.0.00	0.7860	0.7346	0.5798	0.5424	0.5948	0.4084	0.3634	0.3756
ircl_prn.0.10	0.6777	0.6289	0.4154	0.4320	0.4603	0.3139	0.2676	0.2592
ircl_prn.0.20	0.6224	0.5839	0.3678	0.3889	0.3904	0.2789	0.2111	0.2298
ircl_prn.0.30	0.5538	0.5180	0.2995	0.3251	0.3051	0.2114	0.1878	0.1807
ircl_prn.0.40	0.4975	0.4607	0.2674	0.2892	0.2461	0.1870	0.1735	0.1571
ircl_prn.0.50	0.4372	0.4153	0.2213	0.2575	0.2159	0.1549	0.1585	0.1416
ircl_prn.0.60	0.3395	0.3013	0.1589	0.2073	0.1430	0.1095	0.1379	0.1108
ircl_prn.0.70	0.2842	0.2421	0.1287	0.1601	0.1088	0.0907	0.1128	0.0809
ircl_prn.0.80	0.2066	0.1830	0.0773	0.0912	0.0723	0.0650	0.0609	0.0529
ircl_prn.0.90	0.1430	0.1308	0.0404	0.0562	0.0469	0.0396	0.0281	0.0358
ircl_prn.1.00	0.0884	0.0765	0.0230	0.0257	0.0173	0.0222	0.0118	0.0147
P5	0.5520	0.4920	0.3480	0.3640	0.3440	0.2240	0.2040	0.1920
P10	0.4920	0.4520	0.3060	0.3060	0.3080	0.1820	0.2000	0.1960
P15	0.4453	0.4053	0.2720	0.2720	0.2733	0.1693	0.1773	0.1680
P20	0.4040	0.3680	0.2450	0.2520	0.2590	0.1510	0.1570	0.1540
P30	0.3567	0.3260	0.2127	0.2213	0.2240	0.1340	0.1307	0.1293
P100	0.2188	0.1880	0.1164	0.1178	0.1378	0.0808	0.0782	0.0826
P200	0.1447	0.1190	0.0773	0.0847	0.0949	0.0552	0.0554	0.0580
P500	0.0712	0.0592	0.0429	0.0501	0.0523	0.0322	0.0344	0.0349
P1000	0.0390	0.0326	0.0255	0.0291	0.0303	0.0192	0.0226	0.0225

the probability of retrieving the relavent documents. This was evident from the fact that our Hindi-English CLIR system used a very small dictionary of the size of 5,175 words, many of them containing variants of same words. However, the reason for success of this resource in a CLIR task is that, the choice of source language words in the dictionary is motivated by the TFIDF measure of the words from a sufficiently large corpus. Moreover the dictionary creators keyed-in meanings with an IR application in mind, instead of attempting to create an exact synonym dictionary. Also, the restriction on the number of keywords one can type for a give source language word enabled us to capture the homonyms instead of many polysemous variants.

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] 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.

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

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

[7] 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, Edinburgh, Scotland, 2006. ACM Press.

[8] Prasad Pingali, Kula Kekeba Tune, and Vasudeva Varma. Hindi, Telugu, Oromo, English CLIR Evaluation. *Lecture Notes in Computer Science: CLEF 2006 Proceedings*, 2007.

[9] Prasad Pingali and Vasudeva Varma. Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006. In *Working Notes of Cross Language Evaluation Forum 2006*, 2006.

[10] 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.

[11] 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.