

IIIT Hyderabad's CLIR experiments for FIRE-2008

Sethuramalingam S and Vasudeva Varma
Search and Information Extraction Lab
IIIT, Hyderabad, India
sethu@research.iiit.ac.in, vv@iiit.ac.in

ABSTRACT

This paper discourses our CLIR experiments performed for the FIRE¹ workshop. We had submitted our runs for Ad-hoc monolingual document retrieval in Hindi and English, and Ad-hoc cross-lingual document retrieval from Hindi to English, and English to Hindi. In this paper, we describe our English to Hindi and Hindi to English CLIR systems and the experiments conducted on them using the FIRE-2008 dataset. We had used a dictionary-based approach of query translation and transliteration of named entities in the queries using a mapping-based, Compressed Word Format(CWF) algorithm[1]. Disjunctive Query formulation with different scoring weights based on the part of the topic from which they originated gave us overall better performance.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.2 [Database Management]: H.2.3 Languages - Query Languages

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

Cross-Language Information Retrieval, Translation, Transliteration, Evaluation

1. INTRODUCTION

Cross-Language or Cross-Lingual Information Retrieval (CLIR) is the field of Information Retrieval where the user, queries the IR system in one language and retrieves documents in another language. CLIR is gaining impetus among the IR researchers owing to the diversity of languages used in the web, and users' interests towards native language contents.

CLIR provides enough technical challenges to the IR community like (i) How to get the best possible translation for the given query? (ii) How do we do score synonymous translations obtained for a single source query word? (iii) Is there any better ranking scheme than the widely used BM25 algorithm for CLIR?

Current CLIR systems can be broadly classified into two types based on the type of information resources they use.

¹Forum for Information Retrieval Evaluation.
<http://www.isical.ac.in/~clia/>

They are knowledge-based systems and corpus-based systems. Knowledge-based systems use different sources of gathered information such as dictionaries, ontology etc. But the performance of these systems is restrictive. Corpus-based systems may use parallel or comparable corpora which are aligned at word level, sentence level or passage level to learn models automatically. Open research problems in the areas of ranking algorithms, indexing in the cross-lingual perspective are yet to be addressed.

In this paper, we talk about the previous works in CLIR in the following section, the problem statement defined for FIRE is explained in Section 3. The details of our approach are mentioned in Section 4. Our experimentations and analysis are elaborated in Section 5 and Section 6 respectively.

2. RELATED WORK

Research on cross-language retrieval dates back to late 60s, when automatic processing of foreign language documents was added as an extension to the SMART² document retrieval system to German language contents[2]. Multilingual thesaurus was used for the purpose by them. The SIGIR'96 workshop[3] recommended the term "Cross-Language Information Retrieval" for the first time. Many important issues of CLIR like Representation, Resource selection, Merging results, Evaluation, Architecture and other related issues were discussed as part of the workshop. Further refinement on modalities of CLIR was achieved by the 1997 AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. Advancing from there on, the current research on CLIR has expanded to several disciplines like Cross-Language Image Retrieval, Cross-Language Video Retrieval, Cross-Language Geographical Information Retrieval etc.

In the Indian language perspective, not much works were initiated until 2003, when a surprise language exercise was conducted at ACM TALIP³[4]. 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. Following it, in CLEF⁴ 2006, an ad-hoc bi-lingual track in two of the Indian languages namely, Hindi and Telugu was conducted.

²System for the Mechanical Analysis and Retrieval of Text.
<ftp://ftp.cs.cornell.edu/pub/smart/>

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

⁴Cross Language Evaluation Forum.
<http://www.clef-campaign.org/>

There were seven teams participated in the Cross-language track including our team from IIT Hyderabad. We participated and submitted our runs for Hindi and Telugu[5]. Owing to the positive response of the Indian language CLIR in 2006, CLEF launched a separate Indian language subtask and included few more Indian languages like Marathi, Bengali and Tamil in 2007. The number of Indian languages was five in total. The teams were requested to submit one monolingual run in English, one cross-lingual run in Hindi-English and one or more cross-lingual runs in other Indian languages. Five teams[6],[7],[8],[9] including ours participated in the task. This workshop motivated the need for a separate forum like CLEF for the Indian language scenario.

3. PROBLEM STATEMENT

FIRE-2008 consisted of two main tasks namely Ad-hoc monolingual document retrieval in three Indian languages viz. Hindi, Bengali, Marathi and Ad-hoc cross-lingual document retrieval from Indian languages to English and English to Indian languages. The participants were provided with a set of 25 topics in each language for training and 50 topics for testing their CLIR systems.

Each topic contained a unique number identifying the topic, a title, a description and a narration. A title was typically a few words in length and resembled a real world IR query. The description of a topic contained more detailed description of what the user was looking for, as a natural language statement. A narration contained a little more information than the description in the sense that it also gave additional information of what was relevant and what was not. 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 could 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. A sample topic in English is shown in Table 1 below. The system was expected to provide an output of 1000 documents for each topic in a ranked order which were 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 were binary.

We participated in both the Ad-hoc monolingual and the Ad-hoc cross-lingual tasks for Hindi(H) and English(E). We submitted our three best automated runs for all the four combinations of the above two tasks viz. H-E, E-H, E-E and H-H.

4. OUR APPROACH

Our CLIR system was based on a dictionary-based approach of query translation. Dictionaries gathered from different sources were used for this purpose. Named entities found in the queries were transliterated using a mapping-based, Compressed Word Format (CWF) Algorithm for transliteration. Documents were indexed using the Lucene⁵ framework and a vector-based model was used for ranking the documents. Lucene's OKAPI BM25 was used as the similarity metric for scoring the documents.

4.1 Query Processing

⁵<http://lucene.apache.org/>

Table 1: A Sample topic in English

```
<top lang="en">
<num>3</num>
<title>World Cup Cricket 2007</title>
<desc>Find documents reporting 2007 ICC Cricket
World
Cup in West Indies.</desc>
<narr>The documents should contain information re-
garding
the ICC Cricket World Cup which took place in West
Indies.
Reports on the teams which qualified for the semi-finals,
final and eventually the team which won the cup and
India's
performance as a team are of interest.</narr>
</top>
```

Our Query processing module consisted of an n-gram based translation of query words using bi-lingual lexicons, Identification of named entities using Named Entity Recognizers, Transliteration of the identified named entities, and a Boolean query scoring sub-module. A weighted Boolean query would be generated as the output of the Query processing module.

4.1.1 Query Translation

Our Query Translation primary involved translation of query words using bi-lingual lexicons. An n-gram based approach was used to match the entries that have two or more words in the lexicons, thereby increasing the probability of translating multiple-word entries present in the query. We had used Hindi-English lexicons collected from three different sources for translation. An English-Hindi lexicon named 'Shabdanjali' containing 26,633 entries, the Hindi-English WordNet from IIT Bombay, and a manually collected Hindi-English dictionary consisting 6,685 entries.

4.1.2 Named Entities identification

Identification of named entities or Out Of Vocabulary words (OOVs) in the given query is very critical in deciding upon which are the words to be transliterated and which are not. Such a binary classification is of much help rather than recognizing the class of the named entities. In the case of English queries, we used Stanford Named Entity Recognizer to identify the named entities present in the queries. For Hindi queries, the queries are romanized using our custom romanization scheme on prior and then passed as input to our CRF-based Named Entity Recognizing tool. The identified named entities in the queries were passed to the transliteration module for transliteration.

4.1.3 Transliteration of named entities

Our Transliteration module was build on a grapheme-based model, in which transliteration equivalents were identified by mapping the source language names to their equivalents in a target language index, instead of generating them. The basic principle was to compress the source word into its minimal form and align it across an indexed list of target language words to arrive at the top n-equivalents based on the modified edit distance. Mapping-based approach is advantageous over statistical generation-based models for two

main reasons: accuracy and fastness.

For Example, in the case of English to Hindi named entities transliteration, the named entities in English were mapped to their proper romanized Hindi equivalents by comparing their minimal consonant skeletal forms or Compressed Word Formats (CWFs), produced based on a set of linguistic rules. Prior to it, the list of Hindi named entities present in the Hindi corpus had to be prepared. Our CRF-based NER engine was used to identify the named entities present in the whole Hindi corpus. Once we derive the compressed word formats for the English query word and for the list of compressed Hindi words equivalents, we search and match the right equivalent in the index based on our modified Levenshtein algorithm.

To speed up the process of arriving at the right match in the target index, we derived the Compressed Word Format(CWFs) of the target named entities and indexed the CWFs along with their actual forms. When a new query word arrived for transliteration, we generated its CWF form, and compared the CWF of the source query word with the CWFs of the target language entries in the index and the matching entries with modified Levenshtein distance equal to zero were returned. From our experiments, we found that on most of the occasions, the right matches were derived at a modified edit distance of zero during mapping.

4.1.4 Query Scoring

Once the source language queries were translated and transliterated, the resultant target language keywords were used to construct Boolean queries using the OR operator. We used different scores for the words originated from the different parts of the source topic.

Let W_t be the weight assigned to target language words originated from the <title> section of the source topic.

Let W_d be the weight assigned to target language words originated from the description section(<desc>) of the source topic.

Let W_n be the weight assigned to target language words originated from the narration section(<narr>) of the source topic.

Then the ordering of weights can be given as,

$$W_t > W_d > W_n$$

If a particular keyword occurs in multiple sections of the query, it had to be given greater score compared to the other keywords accordingly. Hence, cumulative weight for each word was calculated based on the number of occurrences. Other important keywords like years, numbers, etc. were also given higher weight factors. If t_i be the translated query word in the Boolean query and w_i be the scoring weight associated to it, then the final query output, T for a given source language query from the system would be of the form,

$$T = \bigcup_i w_i.t_i \quad (1)$$

4.2 Indexing and Ranking

Indexing of documents was performed using Lucene's Indexer. In the case of English corpus, stopwords removal and

stemming were performed using Lucene. The Hindi corpus was romanized using our romanization scheme and indexed. For the Hindi corpus, a manually collected list of stopwords containing 246 words was used. For each query, the documents were ranked using Lucene's BM25 algorithm as the similarity metric.

5. EXPERIMENTATIONS

5.1 FIRE Datasets

For both the Ad-hoc monolingual in Hindi task and the Ad-hoc cross-lingual between Hindi and English task, the target document collection in Hindi consisted of 95,215 news articles published in Jagran, a news magazine in Hindi. The target English corpus consisted of around 125,638 news articles from The Telegraph, Calcutta edition, gathered over a period of four years between 2004-2007.

Each document in the target document collection consisted of a unique id mentioned within the <DOCNO> </DOCNO> tags, and the document contents enclosed within the <TEXT></TEXT> tags. The filename of the document was used as the unique id. This document format was found to be maintained across both the Hindi and English document collections. A sample document in English is shown in Table 2.

Table 2: A Sample document in English

```
<DOC>
<DOCNO>1041207_atleisure_index.utf8</DOCNO>
<TEXT>
The Telegraph - Calcutta : At Leisure
Tuesday, December 07, 2004

For Leonardo DiCaprio, it was one of those moments
when he didn't know whether to laugh or cry. There
he was, in a South American rainforest, studying the
effects of mercury poisoning in the Amazon, when he was
confronted by a group of naked Ind...
Six Oscars and big profits for Chicago have sparked a new
frenzy for filming musicals, with Andrew Lloyd Webber's
...
Couples who live together in Britain without getting mar-
ried were warned today that they risk losing their homes
and pos ...
It may come down with a crunch ? or a squelch. Its
creators are hoping for a splash. ...
French fashion legend Pierre Cardin is putting much of
his empire up for sale, seeking $1 billion for his couture
and licens ...

THE FAMILY
Astonishingly, even today, many parents
believe in bringing up their daughters in the security of
a gilded cage, insulated
My first cousin, 27 years old, wants
to get married. Her parents are particularly interested in
a friend of mine who
Actor Diane Kruger at the UK premiere of the film Na-
tional Treasure in London. (Reuters)
</TEXT>
</DOC>
```

5.2 Evaluation

A set of 50 topics in Hindi, Marathi, Bengali and English were provided to the participants by FIRE for evaluation. For our evaluation, we used the test sets in Hindi and English. 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 TREC⁶[10] and CLEF-2007.

We submitted twelve official runs including three official runs for the Ad-hoc English-Hindi and Hindi-English cross-lingual tasks, three mono-lingual runs each in Hindi and English. The submission formats were similar to that of the earlier CLEF formats. The scores of the mono-lingual runs were used as baselines to evaluate the cross-lingual performance of the systems. The details of the runs are given in the Table 3 below.

All the above mentioned runs are generated by combining all the translated and weighted keywords using the Boolean OR operator. The details of the metrics and scores generated from the different runs (mentioned in the Table 3) are shown below. The English-Hindi cross-lingual runs statistics are given in Table 4. The details of Hindi-English cross-lingual runs are mentioned in Table 5. The scores for English-English and Hindi-Hindi mono-lingual runs are specified in Table 6.

6. RUNS ANALYSIS

FIRE team had published their relevance judgements for all the Indian languages and English and had also sent out the scores of the official submissions to each of the participants individually.

As shown in Table 4, Our top MAP score for English-Hindi cross-lingual run was achieved using Title and Description combination. In our best run, our CLIR system was able to retrieve about 53% of total relevant documents for the given 50 test topics. In the monolingual scenario of Hindi-Hindi, our best MAP score was obtained for Title and Narration combination and we were able to retrieve 72% of the total relevant documents. Figure 1 shows the comparison of MAP scores of English-Hindi cross-lingual and Hindi-Hindi monolingual submissions for different submissions. Figure 2 shows the comparison of the cross-lingual and monolingual submissions in terms of precision at different levels of interpolated recall.

For our Hindi-English cross-lingual run, the best MAP score was obtained for Title and Narration combination. Our system was able to retrieve 52% of total relevant documents for the test set of 50 Hindi topics. For the English-English monolingual run, our best MAP score was attained for Title and Narration combination. In the best monolingual run, we were able to retrieve about 91% of the total relevant documents in English. Figure 3 depicts the MAP score comparison of Hindi-English cross-lingual and English-English monolingual run for different runs. Figure 4 presents the comparison of the cross-lingual and monolingual submissions in terms of precision at different levels of interpolated recall.

⁶Text REtrieval Conference. <http://trec.nist.gov/>

Table 3: Runs details

Run Id	Language pair	Description
EHTD	English-Hindi	English queries and Hindi documents. Title and Description fields are used.
EHTN	English-Hindi	English queries and Hindi documents. Title and Narration fields are used.
EHTDN	English-Hindi	English queries and Hindi documents. Title, Description and Narration fields are used.
HETD	Hindi-English	Hindi queries and English documents. Title and Description fields are used.
HETN	Hindi-English	Hindi queries and English documents. Title and Narration fields are used.
HETDN	Hindi-English	Hindi queries and English documents. Title, Description and Narration fields are used.
HHTD	Hindi-Hindi	Hindi Monolingual run. Title and Description fields are used.
HHTN	Hindi-Hindi	Hindi Monolingual run. Title and Narration fields are used.
HHTDN	Hindi-Hindi	Hindi Monolingual run. Title, Description and Narration fields are used.
EETD	English-English	English Monolingual run. Title and Description fields are used.
EETN	English-English	English Monolingual run. Title and Narration fields are used.
EETDN	English-English	English Monolingual run. Title, Description and Narration fields are used.

We found that our English-Hindi cross-lingual performance was 58% of the mono-lingual performance. This can be attributed to factors like exact matching of named entities from English to Hindi on transliteration, and the main dictionary source used for translation viz. "Shabdanjali" was primarily an English to Hindi lexicon with good coverage of English terms. On the other hand, our Hindi-English cross-lingual performance was around 25% of the mono-lingual performance. The reasons for the decrease in performance was lack of linguistic resources like Stemmer or morphological analyzer for Hindi queries, low dictionary coverage on the Hindi side, transliteration mismatches due to the presence of suffixes along with the named entities etc.

7. CONCLUSION AND FUTURE WORK

Our CLIR experiments for FIRE suggest that simple disjunctive query formulation using weighted keywords give an overall better performance in both cross-lingual and monolingual scenarios. But a more complex query formulation involving conjunctions for the most important keywords and

Table 4: English-Hindi cross-lingual Runs details

Metric	EHTD	EHTN	EHTDN
num_q	50	50	50
num_ret	50000	50000	50000
num_rel	3436	3436	3436
num_rel_ret	1708	1820	1762
map	0.1538	0.1516	0.1432
gm_ap	0.0093	0.0229	0.0215
R-prec	0.1687	0.1871	0.1793
bpref	0.1905	0.1918	0.1886
recip_rank	0.3415	0.4677	0.4567
ircl_prn.0.00	0.413	0.5002	0.4961
ircl_prn.0.10	0.3018	0.3257	0.3146
ircl_prn.0.20	0.247	0.2541	0.2491
ircl_prn.0.30	0.1951	0.2039	0.1927
ircl_prn.0.40	0.1672	0.1678	0.1531
ircl_prn.0.50	0.1487	0.1354	0.1208
ircl_prn.0.60	0.1274	0.1023	0.0924
ircl_prn.0.70	0.1076	0.0802	0.077
ircl_prn.0.80	0.0868	0.0645	0.0599
ircl_prn.0.90	0.0471	0.0369	0.0308
ircl_prn.1.00	0.0278	0.0122	0.008
P5	0.26	0.288	0.296
P10	0.258	0.262	0.28
P15	0.2507	0.2627	0.2587
P20	0.233	0.25	0.238
P30	0.22	0.236	0.222
P100	0.1458	0.1702	0.164
P200	0.0973	0.1115	0.1101
P500	0.0564	0.0599	0.059
P1000	0.0342	0.0364	0.0352

negation of non-relevant keywords or information can give us a much higher performance in the cross-lingual scenario. We shall work on such complex query formulations in our future experiments. We also found that in dictionary-based approaches, the coverage of the dictionary, the quality of translations etc. play a significant role in determining the performance of the CLIR system. Moreover in such approaches, the CLIR system performance relies heavily on the lexicons. So, we shall work on other generic approaches of query translation like using Web, aligned parallel corpora, using Wikipedia etc in the future.

8. ACKNOWLEDGEMENTS

Our hearty thanks to our friend, Mr.Vinay Pandit for helping us in building our CLIR systems. Our sincere thanks to Mr. Prasad Pingali for his technical guidance in fine-tuning the system.

9. REFERENCES

[1] Srinivasan C. Janarthanam, Sethuramalingam S, and Udhyakumar Nallasamy. Named entity transliteration for cross-language information retrieval using compressed word format mapping algorithm. In *iNEWS '08: Proceeding of the 2nd ACM workshop on Improving non english web searching*, pages 33–38, New York, NY, USA, 2008. ACM.

[2] G. Salton. Automatic processing of foreign language documents. In *Proceedings of the 1969 conference on*

Table 5: Hindi-English Cross-lingual Runs details

Metric	HETD	HETN	HETDN
num_q	50	50	50
num_ret	50000	50000	50000
num_rel	3779	3779	3779
num_rel_ret	1809	1976	1841
map	0.0907	0.1204	0.1112
gm_ap	0.0197	0.0366	0.0287
R-prec	0.1291	0.1718	0.1541
bpref	0.1408	0.1734	0.1723
recip_rank	0.368	0.4071	0.4108
ircl_prn.0.00	0.399	0.4513	0.435
ircl_prn.0.10	0.2008	0.2761	0.2599
ircl_prn.0.20	0.1586	0.234	0.2105
ircl_prn.0.30	0.119	0.1687	0.1504
ircl_prn.0.40	0.0963	0.1234	0.113
ircl_prn.0.50	0.0775	0.0937	0.0842
ircl_prn.0.60	0.0613	0.0696	0.0601
ircl_prn.0.70	0.0377	0.0417	0.0406
ircl_prn.0.80	0.0271	0.0202	0.021
ircl_prn.0.90	0.0167	0.0083	0.0075
ircl_prn.1.00	0.0004	0.003	0.0024
P5	0.18	0.3	0.268
P10	0.18	0.256	0.266
P15	0.1667	0.2467	0.2387
P20	0.162	0.242	0.226
P30	0.1653	0.2227	0.198
P100	0.1206	0.1514	0.1358
P200	0.0907	0.1105	0.1005
P500	0.0574	0.0642	0.0596
P1000	0.0362	0.0395	0.0368

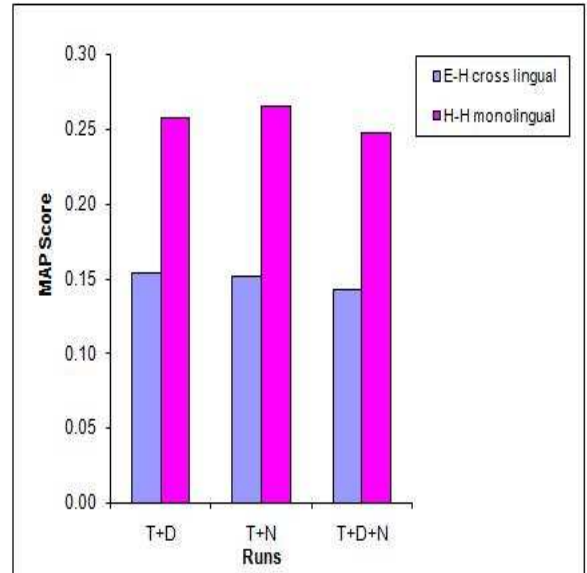


Figure 1: Map score comparison of English-Hindi and Hindi-Hindi submissions

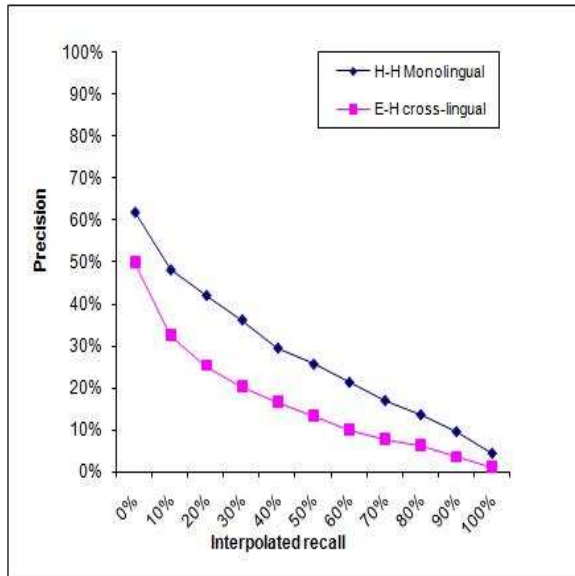


Figure 2: Precision vs Interpolated recall of English-Hindi and Hindi-Hindi submissions

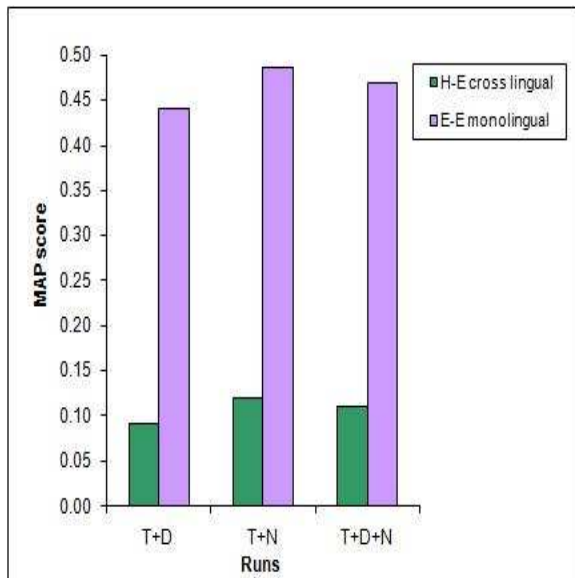


Figure 3: Map score comparison of Hindi-English and English-English submissions

Table 6: Mono-lingual Runs statistics

Run	MAP	GMAP	R-Prec	Bpref
HHTD	0.2579	0.0427	0.2797	0.2964
HHTN	0.2652	0.0534	0.2845	0.3023
HHTDN	0.2472	0.0525	0.2558	0.2773
EETD	0.4416	0.3437	0.4579	0.4889
EETN	0.4863	0.3989	0.4894	0.5218
EETDN	0.469	0.3841	0.4707	0.5167

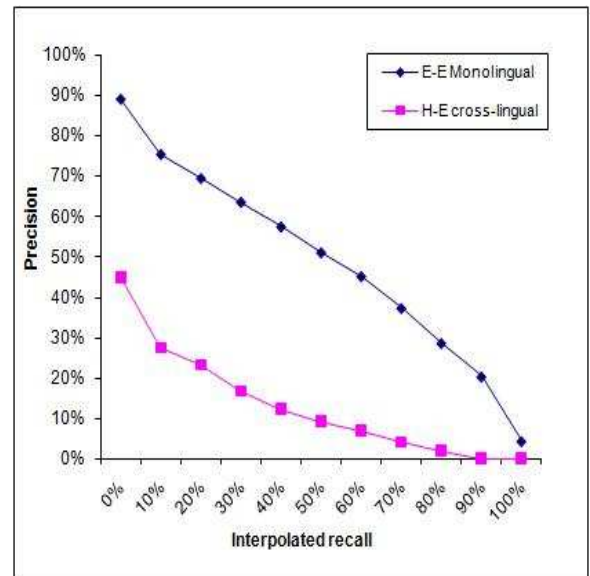


Figure 4: Precision vs Interpolated recall of Hindi-English and English-English submissions

Computational linguistics, pages 1–28, Morristown, NJ, USA, 1969. Association for Computational Linguistics.

- [3] Gregory Grefenstette. Cross-linguistic information retrieval workshop. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, page 344, New York, NY, USA, 1996. ACM.
- [4] Douglas W. Oard. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):79–84, 2003.
- [5] Prasad Pingali and Vasudeva Varma. Hindi and telugu to english cross language information retrieval at clef 2006. *Working Notes of Cross Language Evaluation Forum 2006*, 2006.
- [6] Prasad Pingali and Vasudeva Varma. Iiit hyderabad at clef 2007 - adhoc indian language clir task. *Working Notes of Cross Language Evaluation Forum 2007*, 2007.
- [7] Jagadeesh Jagarlamudi and A Kumaran. Cross-lingual information retrieval system for indian languages. *Working Notes of Cross Language Evaluation Forum 2007*, 2007.
- [8] Debasis Mandal, Sandipan Dandapat, Mayank Gupta, Pratyush Banerjee, and Sudeshna Sarkar. Bengali and hindi to english cross-language text retrieval under limited resources. *Working Notes of Cross Language Evaluation Forum 2007*, 2007.
- [9] Sivaji Bandyopadhyay, Tapabrata Mondal, Sudip Kumar Naskar, Asif Ekbal, Rejwanul Haque, and Srinivasa Rao Godavorthy. Bengali and hindi to english cross-language text retrieval under limited resources. *Working Notes of Cross Language Evaluation Forum 2007*, 2007.
- [10] Ellen M. Voorhees and Donna Harman. Overview of the sixth text retrieval conference (trec-6). *Inf. Process. Manage.*, 36(1):3–35, 2000.