FIRE-2008 at Maryland: English-Hindi CLIR

Tan Xu[1] and Douglas W. Oard[1, 2]
[1]College of Information Studies and
[2]CLIP Lab, Institute for Advanced Computer Studies,
University of Maryland, College Park, MD 20742, USA
tanx@umd.edu, oard@umd.edu

Abstract

In this year's Forum for Information Retrieval Evaluation (FIRE), the University of Maryland participated in the Ad-hoc task cross-language document retrieval task, with English queries and Hindi documents. The experiments focused on evaluating the effectiveness of a “meaning matching” approach based on translation probabilities. The FIRE Hindi test collection provides the first opportunity to carefully assess some of the resources and techniques developed for the Translingual Information Detection, Extraction And Summarization (TIDES) program’s “Surprise Language” exercise in 2003, in which a broad range of language engineering tools were constructed for Hindi in a comparatively short period. The results reported in this paper appear to confirm that some of the language resources developed for the Surprise Language exercise are indeed reusable, and that meaning matching yields reasonably good results with less carefully constructed language resources than had previously been demonstrated.

1. Introduction

In June, 2003, the Defense Advanced Research Projects Agency (DARPA) conducted two “Surprise Language” exercises. The goal of these exercises was to assess the ability of researchers to rapidly apply data-driven techniques for language engineering to an unanticipated language. At the time, DARPA’s Translingual Information, Detection, Extraction, and Summarization (TIDES) program included many of the leading language engineering research groups in the United States of America, but work in that program had to that point focused on just a few languages for which extensive resources had been assembled. A ten-day preliminary exercise for Cebuano confirmed that some progress could be made in a limited time, so the start date for the main exercise was set for June 2 (a Monday). That morning, DARPA announced that Hindi had been selected, with June 30 as the deadline.

The choice of Hindi for the Surprise Language Exercise proved to be considerably more challenging than had been expected. The central unanticipated issue was the diversity of character encodings that were at that time (and, from what we understand, to some extent still are) in common use for Hindi. Ultimately, this problem was overcome by construction of a manually configurable encoding converter. The result of this conversion process was a consistent ASCII transliteration that could easily be used with existing tools. We used this same transliteration for the experiments reported in this paper in order to simplify reuse of resources that had been developed for the Surprise Language exercise.

Sixteen research teams, mostly from the TIDES program, participated in this exercise, developing a broad range of language resources and language engineering systems for Hindi. Some evaluation resources were developed by the National Institute of Standards and Technology (NIST) after the exercise was completed, and we have used one of test collection from that effort as a development evaluation environment for our present experiments. Results from those evaluations were reported by the participants in two issues of the ACM Transactions on Asian Language Information Processing (TALIP) later in 2003 (Oard 2003a; Oard 2003b). That test collection was quite small, however (just 15 topics), and it had been constructed by pooling results from only a few systems. The FIRE evaluation Hindi test collection provides the first opportunity to validate results obtained with the Surprise Language test collection through comparison with results obtained using a larger test collection built using much richer pooling. Assessing the continuing utility of language resources and the information retrieval test collection that had been constructed for the TIDES Surprise Language exercise was therefore the first goal of our experiments.

Much has been learned about Cross-Language Information Retrieval (CLIR) since 2003. The second goal of our experiments was therefore to bring our techniques up to date, applying our best current understanding of effective techniques to ranking Hindi documents based on English queries. Specifically, we applied a technique called Derived Aggregated Meaning Matching (DAMM), which is derived based on the notion of matching what the searcher means with what the author of a document meant. DAMM was designed as an improvement over of Probabilistic Structured Queries (PSQ), which we had used in the Surprise Language exercise. We therefore first compared DAMM with PSQ using the Surprise Language test collection, finding that DAMM indeed seemed to be superior. We therefore used DAMM for our official FIRE submissions, with parameters learned on the Surprise Language test collection.

The remainder of this paper is organized as follows. In the next section, we introduce the DAMM technique for CLIR, comparing it with the earlier PSQ technique. We then describe our preliminary experiments using the Surprise Language test collection, comparing our results with results reported at the time by participants in the Surprise Language exercise. Section 4 describes the three English-Hindi CLIR runs that we submitted for FIRE. Finally, we conclude with some remarks about additional experiments and analysis that would be useful, and other implications of our work.

2. Derived Aggregated Meaning Matching (DAMM)

In general, information retrieval (IR) can be viewed as a task of matching the meaning intended in a query with the meaning expressed in each document, regardless of whether the documents and the queries are expressed using words in the same language (monolingual IR), or in different languages (CLIR). Document independence and term independence assumptions allow us to score each document based on matching the meaning of each query term with the meaning of each corresponding document term. Of course, different terms may share the same meaning. For example, in monolingual IR it is common to treat words that share a common stem as if they express the same meaning. The key insight behind our notion of "meaning matching" (Wang and Oard, 2006) is that we can apply that same perspective fairly directly to CLIR.

A key inspiration for the approach was McCarley experiments with bidirectional translation based on merging retrieval results from each translation direction (McCarley, 1999); our work could be viewed as a more direct way of incorporating of bidirectional evidence in a retrieval model. The basic formulae are a straightforward generalization of Darwish's probabilistic structured queries (PSQ) technique (Darwish and Oard, 2003). In place of the conditional probability *p(f|e)*, we use the notation *p(e* ←→*f)* to indicate the probability that *e* and *f* express the same meaning. It is then straightforward to compute term frequency (*TF*) and document frequency (*DF*) using Equations 1 and 2. Document length normalization is unaffected; it can be performed using only document-language term statistics.

$$TF(e,d_k)=\sum_{f_i} p(e\leftrightarrow f_i)\times TF(f_i,d_k) \qquad (1)$$

$$DF(e)=\sum_{f_i} p(e\leftrightarrow f_i)\times DF(f_i) \qquad (2)$$

Here, *p (e* ←→*f_i)* denotes the probabilities that terms f_i (i = 1, 2, …, n) in document language *F* (in this case, Hindi) share the searcher's intended meaning for the word *e* in query language *E* (in this case, English). If we see a translation f_i appearing one time in document d_k, we can therefore act as if we have seen query term *e* occurring *p(e* ←→*f_i)* times in that document. As pointed by Wang and Oard, comparing with Darwish's PSQ, the fundamental insight behind meaning matching is that there is no need to commit to one translation direction or the other (Wang and Oard, 2006).

In their paper, Wang and Oard then define the computational model for *p (e* ← →*f_i)* in Equation 3:

$$p(e\leftrightarrow f)\approx\sum_{s_j} p(s_j\,|\,e)\times p(s_j\,|\,f) \qquad (3)$$

Where:

- *p(e*←→*f)*: the probability that term e and term f have the same meaning;
- *p(sj|e)*: the probability that term *e* has meaning s_j;
- *p(sj|f)*: the probability that term *f* has meaning s_j.

Oard and Wang found that "synsets" (sets of synonymous terms) could serve as a simple computational model of meaning, and that useful synomym sets could be constructed automatically from statistical word-to-word translation models by looking for words in the same language that were linked by a common translation. Specifically, to find document language synonyms, Equation 4 is used.

$$p(f_j \leftrightarrow f) \approx \sum_{i=1}^{n} p(e_i \mid f) \times p(f_j \mid e_i) \qquad (4)$$

Where $p(f_j \leftrightarrow f)$ is the probability of f_j being a synonym of f.

Since some translations might appear in more than one synset, we need some way of assigning their translation probability across those synset. We have adopted the same greedy method as Wang and Oard was a simple greedy method, iteratively assigning each translation to the synset that would yield the greatest aggregate probability. Specifically, the algorithm works as follows:[1]

1. Compute the aggregate probability that e maps to each s_j: $p(s_j \mid e) = \sum_{f_i \in s_j} p(f_j \mid e)$, and rank all s_j in decreasing order of aggregate probability;
2. Select synset s_j with the largest aggregate probability, remove all of its terms from every synset, and iterate.

To evaluate the effectiveness of this proposed meaning matching model for CLIR, Wang and Oard conducted two sets of experiments: one on ranking French news stories based on English queries, and the other on ranking Chinese news stores based on English queries. Both experiment showed this DAMM method to be quite effective, and also fairly robust across a range of parameter settings, when compared to PSQ (Wang and Oard, 2006). Since we used PSQ in our original Surprise Language experiments, we were interested to see whether DAMM would perform well in the FIRE CLIR task, where our translation probability estimates are far noisier than in the original English-French and English-Chinese experiments.

3. Preliminary Experiments

To evaluate the effectiveness of the DAMM method for CLIR, we first conducted some preliminary experiments, ranking Hindi documents in the Surprise Language test collection based on English queries. This section describes the design of those experiments, the preprocessing that was done for the documents, queries, dictionaries, and statistical synonyms, and generation of the DAMM model.

3.1. Test Collection and IR System

For the CLIR evaluation component of the Surprise Language exercise, 15 TREC-style topic descriptions were created and used to search a collection of 41,697 Hindi documents drawn from several sources. Relatively shallow (top-20) pools were formed from participating systems, and relevance judgments for documents in those pools were created by the Linguistic Data Consortium (LDC). Table 1 shows some statistics of the Surprise Language test collection.

[1] As a result of this method, the summation in Equation 4 will be unused, since greedy aggregation results in unique mappings.

Table 1. Surprise Language test collection statistics.

	Surprise Language Test Collection
Query language	English
Document language	Hindi
Number of topics	15
Number of documents	41,697
Avg. # of rel docs per topic	41

We stripped punctuation from the document collection and removed Hindi terms contained on the stopword list built by University of Massachusetts for the Surprise Language exercise (Larkey, et al, 2003), which consists of 275 Hindi words. We then created a document index based on stemmed Hindi terms. For these preliminary experiments, we only used the statistical stemmer developed by University of California, Berkeley for the Surprise Language exercise (in our submitted runs, we also compared our results with another the YASS stemmer obtained from the FIRE Web site (Majumder, et al, 2006). We formulated queries from the Title, Description and Narrative fields (which we call "TDN" queries). For English queries, we performed pre-translation stopword removal using the English stopword list provided with the InQuery information retrieval system from the University of Massachusetts. For Hindi queries, we performed punctuation removal, stopword removal, and stemming using the same tools that we used for processing the document collection. The Hindi queries serve to establish a useful upper baseline for CLIR effectiveness.

All our experiments were run using the Perl Search Engine (PSE), a document retrieval system based on Okapi BM25 weights that implements PSQ and DAMM. In the Okapi BM25 formula (Robertson and Sparck-Jones, 1997), we used $k_1 = 1.2$, $b = 0.75$, and $k_3 = 7$, as has been commonly used.

3.2. Translation Resources

We used three bilingual lexicons for the development work, one bilingual term list (without translation probabilities) and two statistical translation lexicons. The bilingual term list was constructed from a large bilingual term list assembled by the LDC and a small list of around 400 place names in English and Hindi; both were originally created for the Surprise Language exercise. From this merged bilingual term list, we generated a simple Hindi-to-English translation lexicon (with probabilities) using a uniform distribution over the English translations of each Hindi term (e.g., a Hindi terms with four known English translations would have a translation probability of 0.25 for each). Our second lexicon was based on observed translation counts produced by IBM for the Surprise Language exercise through statistical alignment of parallel English and Hindi news text. After normalization, cleanup, and removal of English and Hindi stop-words and multiword expressions, we constructed a statistical translation lexicon from the remaining counts as follows: word pairs that became identical as a result of normalization were merged (i.e., summing their counts) and $p(h|e)$ (i.e., Hingi given English) probabilities were estimated by dividing the count for a English-Hindi pair by the sum of the counts for all the pairs with the same English term. The count data for our third statistical translation lexicon was contributed by Information Sciences Institute of University of Southern California (USC-ISI). It was preprocessed in the same way as the IBM dictionary. The two statistical translation lexicons that were assembled from counts were combined with weight of 0.6 for the IBM-source lexicon and 0.4 for the LDC-source lexicon to create the Hindi-to-English statistical translation lexicon that was used for the experiments reported in this section (and for our final submissions). Table 2 shows the sizes of these three dictionaries.

Table 2. Size statistics for the translation lexicons.

Source	Number of pairs	English words	Hindi words
LDC	69,195	21,842	33,251
IBM	181,110	50,141	77,517
ISI	512,248	65,366	97,275

Finally, we applied Equation 4 to generate a Hindi synonym lexicon. We pruned the resulting synonym sets by removing any synonym set in which all putative synonyms has probability below .01. This Hindi synonym dictionary consists of 83,522 synsets. Equation 3 and the greedy method was then used to generate a DAMM translation model.

We present the results in three parts: (1) establishing a upper baseline using Hindi queries, (2) establishing a lower baseline using PSQ with English queries, and (3) comparing the retrieval effectiveness of the DAMM with those baselines.

3.3 Monolingual Hindi Baseline

Although monolingual retrieval is not strictly an upper bound for CLIR (because CLIR benefits from expansion effects that might be difficulty to replicate in a monolingual setting), comparing the relative effectiveness of monolingual IR and CLIR is quite common in CLIR evaluation. We obtained monolingual Hindi baseline for the Surprise Language test collection by retrieving documents with TDN queries formulated from topics written in Hindi. This monolingual baseline achieved a Mean Average Precision (MAP) of 0.37. Often, Blind Relevance Feedback (BRF) can improve ranked retrieval results (when averaged over many topics). We therefore automatically expanded the Hindi TDN queries with the top 20 words from the top 20 retrieved documents (based on Okapi weights for each word). Each expansion term was given half the weight of an original (TDN) query term. This resulted in a slight improvement in monolingual MAP to 0.38.

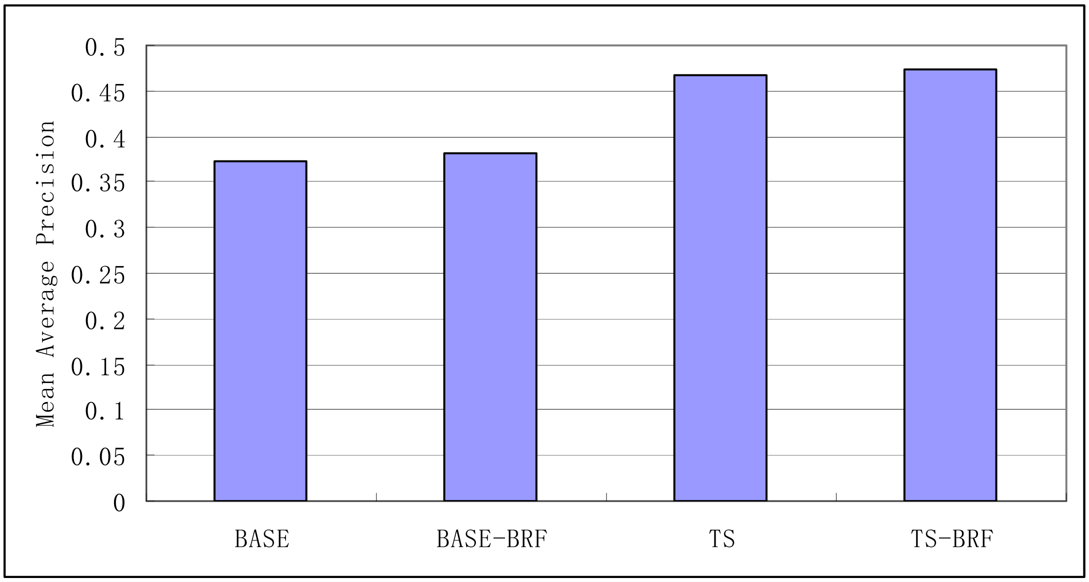

Figure 1. Comparison with the best published Surprise Language exercise results.

Figure 1 shows how these results compare to the best reported monolingual results from the Surprise Language exercise (Larkey, et al, 2003), which we label TS and TS-BRF. It is important to realize that the TS and TS-BRF had the opportunity to contribute to the judgment pools, while our runs did not; with shallow judgment pools that could tend to depress MAP values. It therefore seems reasonable to conclude that our monolingual baselines are at least a credible reference point.

3.4. PSQ Baseline

The best published CLIR results in from the Surprise Language experiments were a MAP of 0.43, which is 91% of the best monolingual MAP. These results were obtained using blind relevance feedback on both the pre-translation English queries and the post-translation Hindi queries. A somewhat lower MAP, 0.41, which is 86% of the best monolingual results, was reported when BRF was used only for pre-translation expansion of the English queries. Table 3 summarizes these results.

Table 3. Best CLIR effectiveness measures, Surprise Language test collection, 15 queries.

Task	Expansion	MAP	Top 5	Top 10	Top 20
CLIR	English only	.4067	3.6	6.3	10.6
CLIR	English+Hindi	.4298	3.7	6.5	11.0

We chose PSQ as a lower baseline for our CLIR experiments because at the time of the Surprise Language exercise it yielded among the best retrieval effectiveness results, and because we had actually used PSQ for our own Surprise Language exercise experiments. Table 5 shows the results for our PSQ baseline run, in which we did not use BRF in either language. We obtained a MAP of 0.29 for this configuration, which is 76% of our monolingual baseline. Table 5 shows the results, which again are considerably below the best published results (perhaps because of the shallow pooling).

Table 4. PSQ effectiveness measures, Surprise Language test collection, 15 queries.

Task	Expansion	MAP	Top 5	Top 10	Top 20
CLIR	None	.2900	2.7	4.8	8.5

3.5. DAMM

Translation probability estimates are less reliable for rare events, so it is common to remove rare translations from such lexicons as a pre-processing step and then renormalize the translation probabilities. In our work, we use apply a threshold on the cumulative probability (i.e., we sum the probabilities from largest to smallest, stopping when the sum reaches our pre-set threshold). Figure 2 illustrates this effect; the monolingual Hindi baseline was used to set the 100% value on the vertical axis in that Figure. As Figure 2 shows, the DAMM MAP increases monotonically, especially after a CPT of 0.4, and the highest MAP, which is about 82.5% of monolingual IR MAP, is achieved when CPT is set to 1.0, which means all the translation alternatives are included. This is a reasonable phenomenon, since the more translations are involved, the larger effect aggregation in DAMM is likely to have, which in turn could improve the effectiveness of meaning matching.

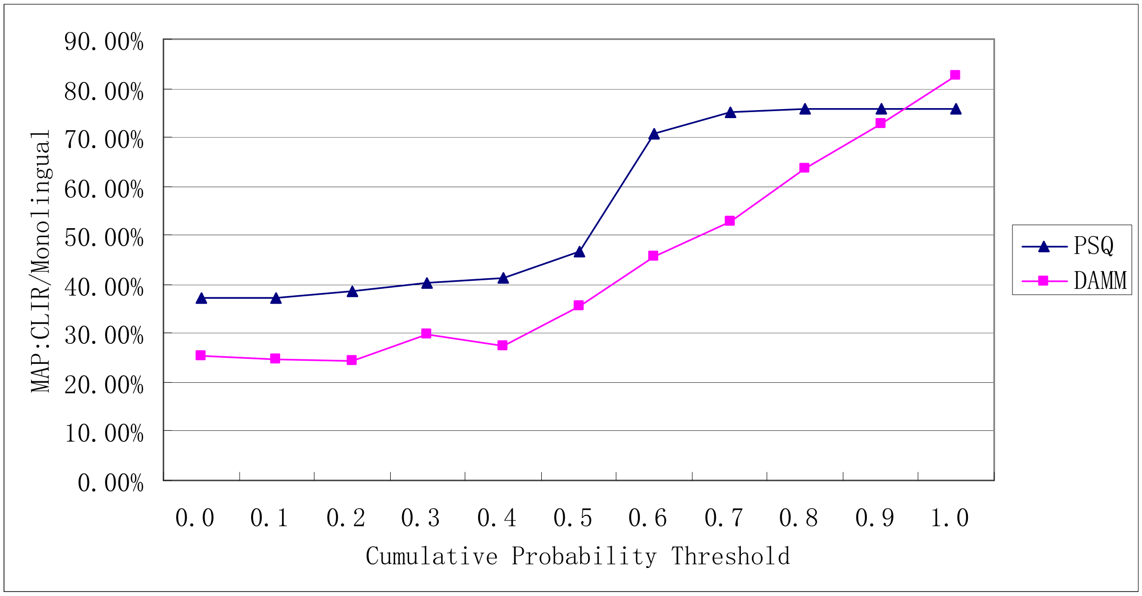

Figure 2. Sensitivity of DAMM and PSQ to cumulative probability threshold.

Comparing its effectiveness with PSQ, although only at high CPT regions seems to indicate some advantage of using DAMM, the best results for DAMM look to be somewhat better (with a 7% relative improvement that we did not test for statistical significance because of the small number of topics). Also, from the figure, we noticed that at low CPT, both methods gave a bad performance, and the DAMM even worse, only after more relatively lower probability translations were involved, the effectiveness of the both methods increased dramatically. This is an indication of that the language resources we are using are not well constructed, since one of the most possible reason for this result is that some spurious translation alternatives are gives better estimation of translation probability. But, even with these less carefully constructed language resources, DAMM could achieve an 82.5% of monolingual IR MAP. Thus, we conclude that it is a reasonable choice for CLIR task, and that 1.0 is a useful threshold. That, then, is the configuration that we used for the experiments reported in the next section.

4. FIRE-2008 Experiments

We participated only in the CLIR task, using English queries to search Hindi documents. In this section we describe the test collection, our submitted runs, and our results.

4.1. Test Collection

Some statistics for the FIRE 2008 Hindi and English test collection are listed in Table 5.

Table 5. FIRE-2008 Test Collection Statistics.

	FIRE-2008 Test Collection	
Query language	English	English
Document language	Hindi	English
Number of search topics	50	50
Number of documents	95,215	125,586
Avg. # of rel docs per topic	68	75

We used the English documents for pre-translation expansion and the Hindi documents only as the search target (i.e., we did not conduct post-translation expansion for our submitted runs). Our preprocessing steps were unchanged form the preliminary experiments described above with one exception: in addition to the Berkeley stemmer, we also tried the YASS stemmer.

4.2. Submitted Runs

We submitted three runs for official scoring in the CLIR task:

1. clir-EH-umd-man0: DAMM with CPT of 1.0, using the Berkeley stemmer, no BRF
2. clir-EH-umd-man1: DAMM with CPT of 1.0, using YASS stemmer, no BRF
   The parameters set for YASS are: d6 distance for clustering; matrix_cut = 4; clustering threshold = 1.5 (Majumder, et al, 2006).
3. clir-EH-umd-man2: DAMM with CPT of 1.0, using the Berkeley stemmer, pre-translation (English) BRF
   When expanded the English TDN queries through English document collection, we selected the top 20 words from the top 20 retrieved documents (based on Okapi weights for each word). Each expansion term was given half the weight of an original (TDN) query term.

4.3. Results

Figure 3 reports our three run results for the MAP. For comparison, we also show the mean MAP across all CLIR runs, and best Hindi monolingual IR MAP. The result (clir-EH-umd-man1) produced by using YASS stemmer turns out to be better than the result using Berkeley stemmer (clir-EH-umd-man0). A Wilcoxon signed rank test shows YASS significant outperformed Berkeley stemmer on all topics. Pre-translation BRF on English queries, unsurprisingly improves the result based on Berkeley stemmer by 0.06 in MAP.

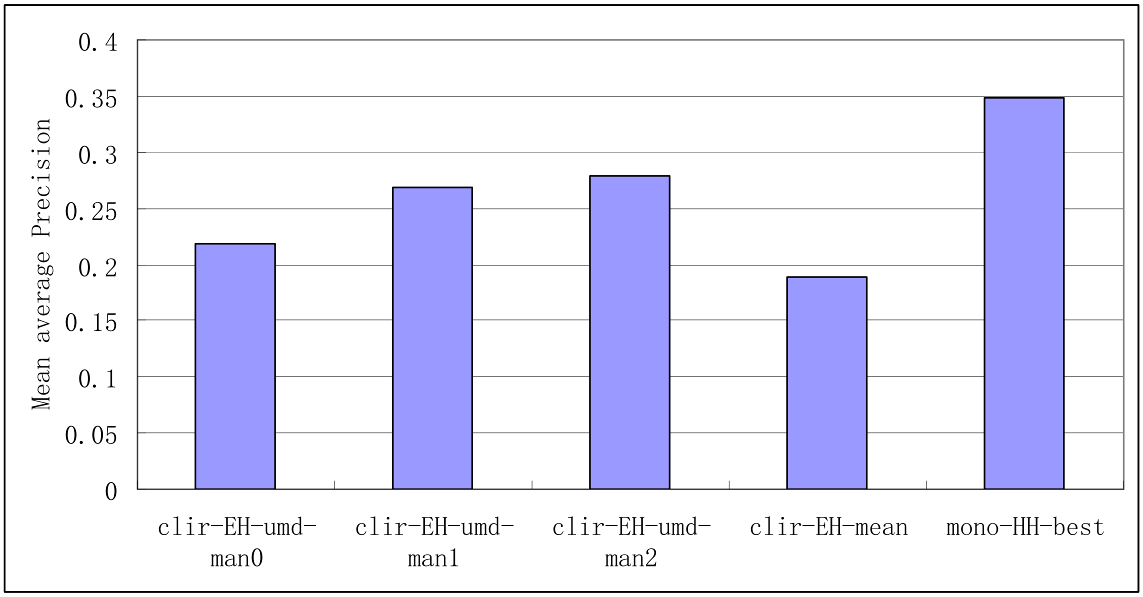

Figure 3 FIRE'08 Official Results

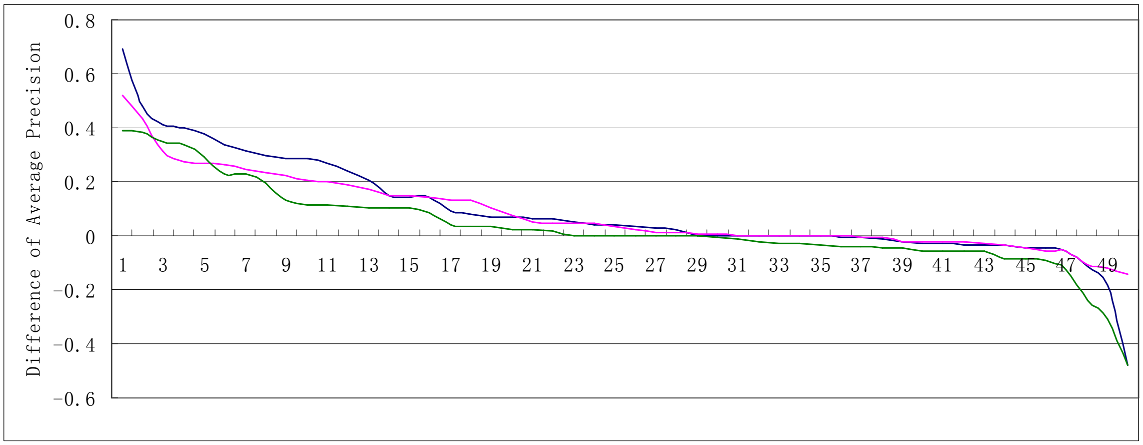

Figure 4 Comparison on each queries

Although our runs statistical significantly outperformed the median of all the submission CLIR runs, and our best result achieved 75% of the best Hindi monolingual run. We want to further investigate what actually happened through query-by-query comparison. We plot the non-interpolated average precision (AP) difference for each query between our three runs and the mean AP among all the CLIR submissions, see Figure 4. Among the 50 queries, our best run (clir-EH-umd-man2) has 29 higher AP, 16 lower and 5 equal. If we do not consider queries that have less than 4 relevant documents (9 queries are ignored in total), clir-EH-umd-man2 performs better on 68.29% queries than average, as shown in Figure 5.

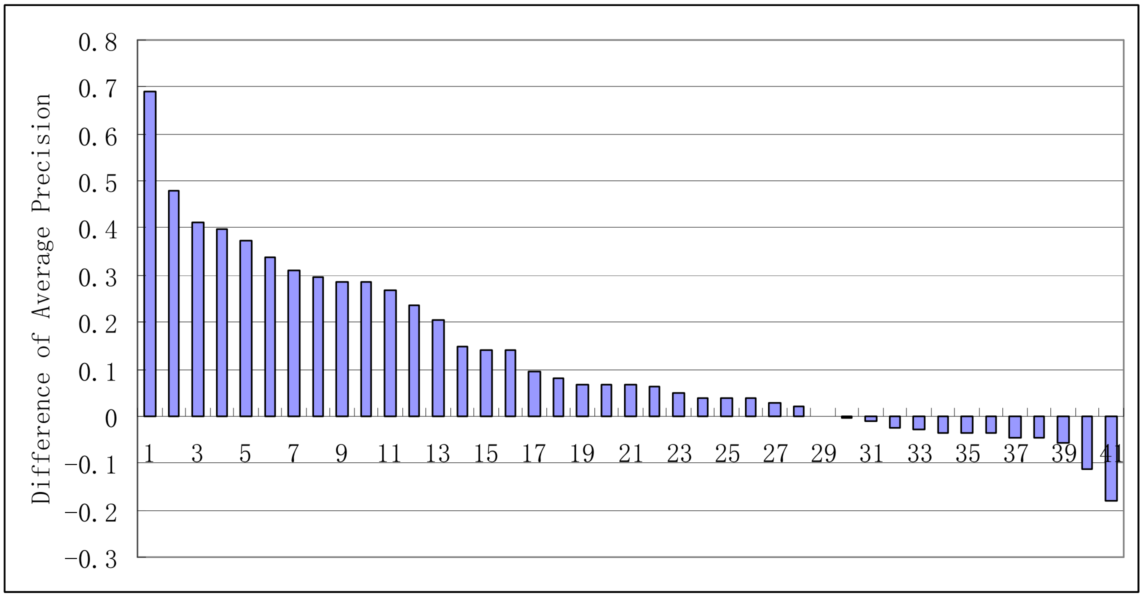

Figure 5 Comparison on Queries with Relevant Documents larger than 4

5. Conclusions and Future Work

Our FIRE 2008 experiments provide some evidence that DAMM, which using bidirectional translation knowledge together with statistical synonymy, gives good performance when dealing with CLIR task, ranking Hindi documents according to English queries specifically. And our preliminary experiments show that comparing with PSQ, the effectiveness of DAMM is also acceptable, especially when high CPT is adopted, which in turn involved more translation alternatives, DAMM outperformed PSQ with significant improvement. Despite the poor language resources and small training collection we are using, our experiments somewhat confirm this finding. This novel technique of using statistical translation knowledge for searching information across language boundaries has been proven to be effective.

However, several things should be considered for improving and further more careful designed experiments are required to be conducted. First, an obvious limitation of our current experiments is the poor quality of our language resources, especially the Hindi-English translation lexicon, for example, an addition experiment on the Surprise test collection by using this lexicon only for CLIR, only gives a MAP of 0.1443. Since this is an important resource when computing the DAMM statistical translation model, we need to build a stronger Hindi-English probability translation lexicon. Secondly, decisions for some parameter settings in our experiments were arbitrary, e.g., synonyms were cut off at the probability of 0.1, and the selections of BRF and YASS parameters. In the future, we plan to explore a broader spectrum of parameter settings, which will hopefully provide us better results. Thirdly, according to our official runs, YASS performs better than Berkeley stemmer, which we are going to further study its effectiveness in our other experiments. Fourthly, a post-translation Hindi query expansion would be conducted, which often will improve the system effectiveness. Last but not least, the DAMM method only tried the greedy method of aggregation. Thus, it may also be worth comparing other techniques that assign each translation alternative to multiple synsets with some weighting factor, e.g., based on information such as orthographic similarity between the translation and words in each synset.

Reference

Darwish, K. and Oard, D.W. (2003). Probabilistic structured query methods. In Proceedings of the 21st Annual 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.338-344. ACM Press.

He, D., Oard, D.W., and et al. (2003). Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi. ACM Transactions on Asian Language Information Processing, 2(3), 219-244.

Larkey, L.S., Connell, M.E., and Abduljaleel N. (2003). Hindi CLIR in Thirty Days. ACM Transactions on Asain Language Information Processing. 2(2), pp. 130-142.

Majumder, P., Mitra, M. and et al. (2006). YASS: Yet Another Suffix Stripper. ACM Trans. Inf. Syst. 25(4), 18.

McCarley, J.S. (1999). Should we translate the documents or the queries in cross-language information retrieval? In Proceedings of the 37th Annual Conference of the Association for Computational Linguistics, pp.208-214.

Oard, D.W. (2003). The Surprise Langauge Exercises. ACM Transactions on Asian Language Information Processing, 2(3)79-84.

Robertson, S.E. and Sparck-Jones, K. (1997). Simple proven approaches to text retrieval. Cambridge University Computer Laboratory.

Wang, J. and Oard, D.W. (2006). Combining Bidirectional Transation and Synonymy for Cross-Language Information Retrieval. SIGIR'06, August 6-11, Seattle, Washington, USA.