

Cross-lingual Information Retrieval using Hidden Markov Models

Jinxi Xu
BBN Technologies
70 Fawcett St.
Cambridge, MA, USA 02138
jxu@bbn.com

Ralph Weischedel
BBN Technologies
70 Fawcett St.
Cambridge, MA, USA 02138
weischedel@bbn.com

Abstract

This paper presents empirical results in cross-lingual information retrieval using English queries to access Chinese documents (TREC-5 and TREC-6) and Spanish documents (TREC-4). Since our interest is in languages where resources may be minimal, we use an integrated probabilistic model that requires only a bilingual dictionary as a resource. We explore how a combined probability model of term translation and retrieval can reduce the effect of translation ambiguity. In addition, we estimate an upper bound on performance, if translation ambiguity were a solved problem. We also measure performance as a function of bilingual dictionary size.

1 Introduction

Cross-language information retrieval (CLIR) can serve both those users with a smattering of knowledge of other languages and also those fluent in them. For those with limited knowledge of the other language(s), CLIR offers a wide pool of documents, even though the user does not have the skill to prepare a high quality query in the other language(s). Once documents are retrieved, machine translation or human translation, if desired, can make the documents usable. For the user who is fluent in two or more languages, even though he/she may be able to formulate good queries in each of the source languages, CLIR relieves the user from having to do so.

Most CLIR studies have been based on a variant of tf-idf; our experiments instead use a hidden Markov model (HMM) to estimate the

probability that a document is relevant given the query. We integrated two simple estimates of term translation probability into the mono-lingual HMM model, giving an estimate of the probability that a document is relevant given a query in another language.

In this paper we address the following questions:

- How can a combined probability model of term translation and retrieval minimize the effect of translation ambiguity? (Sections 3, 5, 6, 7, and 10)
- What is the upper bound performance using bilingual dictionary lookup for term translation? (Section 8)
- How much does performance degrade due to omissions from the bilingual dictionary and how does performance vary with size of such a dictionary? (Sections 8-9)

All experiments were performed using a common baseline, an HMM-based (mono-lingual) indexing and retrieval engine. In order to design controlled experiments for the questions above, the IR system was run without sophisticated query expansion techniques.

Our experiments are based on the Chinese materials of TREC-5 and TREC-6 and the Spanish materials of TREC-4.

2 HMM for Mono-Lingual Retrieval

Following Miller et al., 1999, the IR system ranks documents according to the probability that a document D is relevant given the query Q , $P(D \text{ is } R | Q)$. Using Bayes Rule, and the fact that $P(Q)$ is constant for a given query, and our initial assumption of a uniform a priori

probability that a document is relevant, ranking documents according to $P(Q|D \text{ is } R)$ is the same as ranking them according to $P(D \text{ is } R|Q)$. The approach therefore estimates the probability that a query Q is generated, given the document D is relevant. (A glossary of symbols used appears below.)

We use x to represent the language (e.g. English) for which retrieval is carried out. According to that model of monolingual retrieval, it can be shown that

$$p(Q | D \text{ is } R) = \prod_{W \text{ in } Q} (aP(W | G_x) + (1-a)P(W | D)),$$

where W 's are query words in Q . Miller et al. estimated probabilities as follows:

- The transition probability a is 0.7 using the EM algorithm (Rabiner, 1989) on the TREC4 ad-hoc query set.
- $P(W | G_x) = \frac{\text{number of occurrences of } W \text{ in } C_x}{\text{length of } C_x}$

which is the general language probability for word W in language x .

- $P(W | D) = \frac{\text{number of occurrences of } W \text{ in } D}{\text{length of } D}$

In principle, any large corpus C_x that is representative of language x can be used in computing the general language probabilities. In practice, the collection to be searched is used for that purpose. The length of a

Q	a query
Q_x	English query
D	a document
D_y	a document in foreign language y
D is R	document is relevant
W	a word
G_x	an English corpus
C_x	a corpus in language x
W_x	an English word _____ W_y a word in foreign language y
BL	a bilingual dictionary

A Glossary of Notation used in Formulas

collection is the sum of the document lengths.

3 HMM for Cross-lingual IR

For CLIR we extend the query generation process so that a document D_y written in language y can generate a query Q_x in language x . We use W_x to denote a word in x and W_y to denote a word in y . As before, to model general query words from language x , we estimate $P(W_x | G_x)$ by using a large corpus C_x in language x . Also as before, we estimate $P(W_y | D_y)$ to be the sample distribution of W_y in D_y .

We use $P(W_x | W_y)$ to denote the probability that W_y is translated as W_x . Though terms often should not be translated independent of their context, we make that simplifying assumption here. We assume that the possible translations are specified by a bilingual lexicon BL. Since the event spaces for W_y 's in $P(W_y | D_y)$ are mutually exclusive, we can compute the output probability $P(W_x | D_y)$:

$$P(W_x | D_y) = \sum_{W_y \text{ in } BL} P(W_y | D_y) P(W_x | W_y)$$

We compute $P(Q_x | D_y \text{ is } R)$ as below:

$$P(Q_x | D_y \text{ is } R) = \prod_{W_x \text{ in } Q_x} (aP(W_x | G_x) + (1-a)P(W_x | D_y))$$

The above model generates queries from documents, that is, it attempts to determine how likely a particular query is given a relevant document. The retrieval system, however, can use either query translation or document translation. We chose query translation over document translation for its flexibility, since it allowed us to experiment with a new method of estimating the translation probabilities without changing the index structure.

4 Experimental Set-up

For retrieval using English queries to search Chinese documents, we used the TREC5 and TREC6 Chinese data which consists of 164,789 documents from the Xinhua News Agency and People's Daily, averaging 450 Chinese characters/document. Each of the TREC topics has three Chinese fields: title, description and

narrative, plus manually translated, English versions of each. We corrected some of the English queries that contained errors, such as “Dali Lama” instead of the correct “Dalai Lama” and “Medina” instead of “Medellin.” Stop words and stop phrases were removed. We created three versions of Chinese queries and three versions of English queries: short (title only), medium (title and description), and long (all three fields).

For retrieval using English queries to search Spanish documents, we used the TREC4 Spanish data, which has 57,868 documents. It has 25 queries in Spanish with manual translations to English. We will denote the Chinese data sets as Trec5C and Trec6C and the Spanish data set as Trec4S.

We used a Chinese-English lexicon from the Linguistic Data Consortium (LDC). We pre-processed the dictionary as follows:

1. Stem Chinese words via a simple algorithm to remove common suffixes and prefixes.
2. Use the Porter stemmer on English words.
3. Split English phrases into words. If an English phrase is a translation for a Chinese word, each word in the phrase is taken as a separate translation for the Chinese word.¹
4. Estimate the translation probabilities. (We first report results assuming a uniform distribution on a word’s translations. If a Chinese word c has n translations e_1, e_2, \dots, e_n , each of them will be assigned equal probability, i.e., $P(e_i|c) = 1/n$. Section 10 supplements this with a corpus-based distribution.)
5. Invert the lexicon to make it an English-Chinese lexicon. That is, for each English word e , we associate it with a list of Chinese words c_1, c_2, \dots, c_m together with non-zero translation probabilities $P(e|c_i)$.

The resulting English-Chinese lexicon has 80,000 English words. On average, each English word has 2.3 Chinese translations.

For Spanish, we downloaded a bilingual English-Spanish lexicon from the Internet (<http://www.activa.arrakis.es>) containing around 22,000 English words (16,000 English stems) and processed it similarly. Each English word has around 1.5 translations on average. A co-occurrence based stemmer (Xu and Croft, 1998) was used to stem Spanish words. One difference from the treatment of Chinese is to include the English word as one of its own translations in addition to its Spanish translations in the lexicon. This is useful for translating proper nouns, which often have identical spellings in English and Spanish but are routinely excluded from a lexicon.

One problem is the segmentation of Chinese text, since Chinese has no spaces between words. In these initial experiments, we relied on a simple sub-string matching algorithm to extract words from Chinese text. To extract words from a string of Chinese characters, the algorithm examines any sub-string of length 2 or greater and recognizes it as a Chinese word if it is in a predefined dictionary (the LDC lexicon in our case). In addition, any single character which is not part of any recognized Chinese words in the first step is taken as a Chinese word. Note that this algorithm can extract a compound Chinese word as well as its components. For example, the Chinese word for “particle physics” as well as the Chinese words for “particle” and “physics” will be extracted. This seems desirable because it ensures the retrieval algorithm will match both the compound words as well as their components. The above algorithm was used in processing Chinese documents and Chinese queries.

English data from the 2 GB of TREC disks 1&2 was used to estimate $P(W|G_{English})$, the general language probabilities for English words. The evaluation metric used in this study is the average precision using the trec_eval program (Voorhees and Harman, 1997). Mono-lingual retrieval results (using the Chinese and Spanish queries) provided our baseline, with the HMM retrieval system (Miller et al, 1999).

¹ Clearly, this is not correct; however, it simplified implementation.

5 Retrieval Results

Table 2 reports average precision for mono-lingual retrieval, average precision for cross-lingual, and the relative performance ratio of cross-lingual retrieval to mono-lingual.

Relative performance of cross-lingual IR varies between 67% and 84% of mono-lingual IR. Trec6 Chinese queries have a somewhat higher relative performance than Trec5 Chinese queries. Longer queries have higher relative performance than short queries in general. Overall, cross-lingual performance using our HMM retrieval model is around 76% of mono-lingual retrieval. A comparison of our mono-lingual results with Trec5 Chinese and Trec6 Chinese results published in the TREC proceedings (Voorhees and Harman, 1997, 1998) shows that our mono-lingual results are close to the top performers in the TREC conferences. Our Spanish mono-lingual performance is also comparable to the top automatic runs of the TREC4 Spanish task (Harman, 1996). Since these mono-lingual results were obtained without using sophisticated query processing techniques such as query expansion, we believe the mono-lingual results form a valid baseline.

Query sets	Mono-lingual	Cross-lingual	% of Mono-lingual
Trec5C-short	0.2830	0.1889	67%
Trec5C-medium	0.3427	0.2449	72%
Trec5C-long	0.3750	0.2735	73%
Trec6C-short	0.3423	0.2617	77%
Trec6C-medium	0.4606	0.3872	84%
Trec6C-long	0.5104	0.4206	82%
Trec4S	0.2252	0.1729	77%

Table 2: Comparing mono-lingual and cross-lingual retrieval performance. *The scores on the monolingual and cross-lingual columns are average precision.*

6 Comparison with other Methods

In this section we compare our approach with two other approaches. One approach is "simple

substitution", i.e., replacing a query term with all its translations and treating the translated query as a bag of words in mono-lingual retrieval. Suppose we have a simple query $Q=(a, b)$, the translations for a are a_1, a_2, a_3 , and the translations for b are b_1, b_2 . The translated query would be $(a_1, a_2, a_3, b_1, b_2)$. Since all terms are treated as equal in the translated query, this gives terms with more translations (potentially the more common terms) more credit in retrieval, even though such terms should potentially be given less credit if they are more common. Also, a document matching different translations of one term in the original query may be ranked higher than a document that matches translations of different terms in the original query. That is, a document that contains terms a_1, a_2 and a_3 may be ranked higher than a document which contains terms a_1 and b_1 . However, the second document is more likely to be relevant since correct translations of the query terms are more likely to co-occur (Ballesteros and Croft, 1998).

A second method is to structure the translated query, separating the translations for one term from translations for other terms. This approach limits how much credit the retrieval algorithm can give to a single term in the original query and prevents the translations of one or a few terms from swamping the whole query. There are several variations of such a method (Ballesteros and Croft, 1998; Pirkola, 1998; Hull 1997). One such method is to treat different translations of the same term as synonyms. Ballesteros, for example, used the INQUERY (Callan et al, 1995) synonym operator to group translations of different query terms. However, if a term has two translations in the target language, it will treat them as equal even though one of them is more likely to be the correct translation than the other. By contrast, our HMM approach supports translation probabilities. The synonym approach is equivalent to changing all non-zero translation probabilities $P(W_x|W_y)$'s to 1 in our retrieval function. Even estimating uniform translation probabilities gives higher weights to unambiguous translations and lower weights to highly ambiguous translations.

These intuitions are supported empirically by the results in Table 3. We can see that the HMM performs best for every query set. Simple substitution performs worst. The synonym approach is significantly better than substitution, but is consistently worse than the HMM

	Substitution	Synonym	HMM
Trec5C-long	0.0391	0.2306	0.2735
Trec6C-long	0.0941	0.3842	0.4206
Trec4S	0.0935	0.1594	0.1729

Table 3: Comparing different methods of query translation. *All numbers are average precision.*

7 Impact of Translation Ambiguity

To get an upper bound on performance of any disambiguation technique, we manually disambiguated the Trec5C-medium, Trec6C-medium and Trec4S queries. That is, for each English query term, a native Chinese or Spanish speaker scanned the list of translations in the bilingual lexicon and kept one translation deemed to be the best for the English term and discarded the rest. If none of the translations was correct, the first one was chosen.

The results in Table 4 show that manual disambiguation improves performance by 17% on Trec5C, 4% on Trec4S, but not at all on Trec6C. Furthermore, the improvement on Trec5C appears to be caused by big improvements for a small number of queries. The one-sided t-test (Hull, 1993) at significance level 0.05 indicated that the improvement on Trec5C is not statistically significant.

It seems surprising that disambiguation does not help at all for Trec6C. We found that many terms have more than one valid translation. For example, the word “flood” (as in “flood control”) has 4 valid Chinese translations. Using all of them achieves the desirable effect of query expansion. It appears that for Trec6C, the benefit of disambiguation is cancelled by choosing only one of several alternatives, discarding those other good translations. If multiple correct

translations were kept in disambiguation, the improvement would be 4% for Trec6C-medium.

The results of this manual disambiguation suggest that there are limits to automatic disambiguation.

Query sets	Degree of Disambiguation		
	None	Manual	% of Mono-lingual
Trec5C-medium	0.2449	0.2873 (+17%)	84%
Trec6C-medium	0.3872	0.3830 (-1%)	83%
Trec4S	0.1729	0.1799 (+4%)	80%

Table 4: The effect of disambiguation on retrieval performance. *The scores reported are average precision.*

8 Impact of Missing Translations

Results in the previous section showed that manual disambiguation can bring performance of cross-lingual IR to around 82% of mono-lingual IR. The remaining performance gap between mono-lingual and cross-lingual IR is likely to be caused by the incompleteness of the bilingual lexicon used for query translation, i.e., missing translations for some query terms. This may be a more serious problem for cross-lingual IR than ambiguity. To test the conjecture, for each English query term, a native speaker in Chinese or Spanish manually checked whether the bilingual lexicon contains a correct translation for the term in the context of the query. If it does not, a correct translation for the term was added to the lexicon. For the query sets Trec5C-medium and Trec6C-medium, there are 100 query terms for which the lexicon does not have a correct translation. This represents 19% of the 520 query terms (a term is counted only once in one query). For the query set Trec4S, the percentage is 12%.

The results in Table 5 show that with augmented lexicons, performance of cross-lingual IR is 91%, 99% and 95% of mono-lingual IR on Trec5C-medium, Trec6C-medium and Trec4S.

The improvement over using the original lexicon is 28%, 18% and 23% respectively. The results demonstrate the importance of a complete lexicon. Compared with the results in section 7, the results here suggest that missing translations have a much larger impact on cross-lingual IR than translation ambiguity does.

Query sets	Original lexicon	Augmented lexicon	% of Mono-lingual
Trec5C-medium	0.2449	0.3131 (+28%)	91%
Trec6C-medium	0.3872	0.4589 (+18%)	99%
Trec4S	0.1729	0.2128 (+23%)	95%

Table 5: The impact of missing the right translations on retrieval performance. All scores are average precision.

9 Impact of Lexicon Size

In this section we measure CLIR performance as a function of lexicon size. We sorted the English words from TREC disks 1&2 in order of decreasing frequency. For a lexicon of size n , we keep only the n most frequent English words.

The upper graph in Figure 1 shows the curve of cross-lingual IR performance as a function of the size of the lexicon based on the Chinese short and medium-length queries. Retrieval performance was averaged over Trec5C and Trec6C. Initially retrieval performance increases sharply with lexicon size. After the dictionary exceeds 20,000, performance levels off. An examination of the translated queries shows that words not appearing in the 20,000-word lexicon usually do not appear in the larger lexicons either. Thus, increases in the general lexicon beyond 20,000 words did not result in a substantial increase in the coverage of the query terms.

The lower graph in Figure 1 plots the retrieval performance as a function of the percent of the full lexicon. The figure shows that short queries are more susceptible to incompleteness of the

lexicon than longer queries. Using a 7,000-word lexicon, the short queries only achieve 75% of their performance with the full lexicon. In comparison, the medium-length queries achieve 87% of their performance.

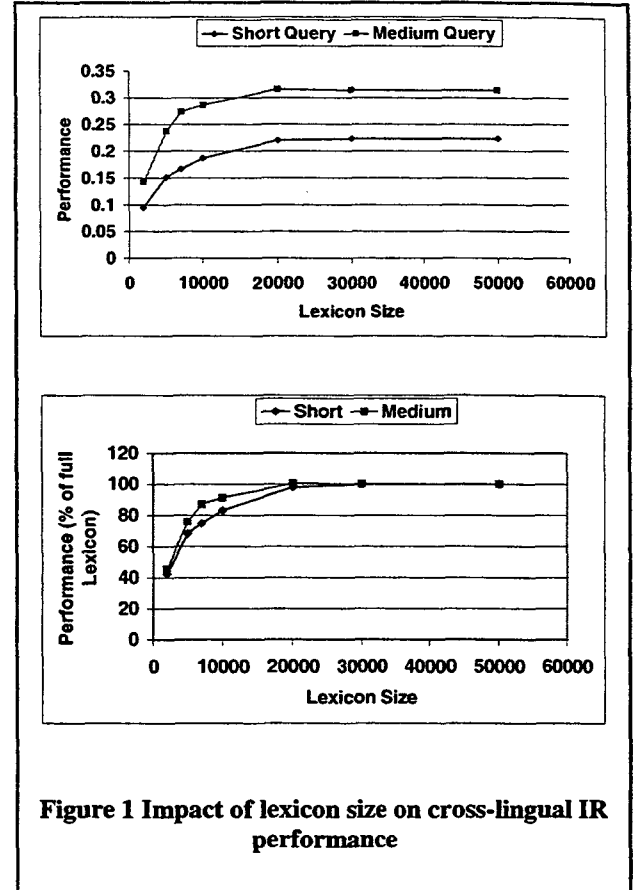


Figure 1 Impact of lexicon size on cross-lingual IR performance

We categorized the missing terms and found that most of them are proper nouns (especially locations and person names), highly technical terms, or numbers. Such words understandably do not normally appear in traditional lexicons. Translation of numbers can be solved using simple rules. Transliteration, a technique that guesses the likely translations of a word based on pronunciation, can be readily used in translating proper nouns.

Another technique is automatic discovery of translations from parallel or non-parallel corpora (Fung and Mckeown, 1997). Since traditional lexicons are more or less static repositories of knowledge, techniques that discover translation from newly published materials can supplement them with corpus-specific vocabularies.

10 Using a Parallel Corpus

In this section we estimate translation probabilities from a parallel corpus rather than assuming uniform likelihood as in section 4. A Hong Kong News corpus obtained from the Linguistic Data Consortium has 9,769 news stories in Chinese with English translations. It has 3.4 million English words. Since the documents are not exact translations of each other, occasionally having extra or missing sentences, we used document-level co-occurrence to estimate translation probabilities. The Chinese documents were “segmented” using the technique discussed in section 4. Let $co(e, c)$ be the number of parallel documents where an English word e and a Chinese word c co-occur, and $df(c)$ be the document frequency of c . If a Chinese word c has n possible translations e_1 to e_n in the bilingual lexicon, we estimate the corpus translation probability as:

$$P_{\text{corpus}}(e_i|c) = \frac{co(e_i, c)}{\text{MAX}(df(c), \sum_{i=1}^{i=n} co(e_i, c))}$$

Since several translations for c may co-occur in a document, $\sum co(e_i, c)$ can be greater than $df(c)$. Using the maximum of the two ensures that

$$\sum P_{\text{corpus}}(e_i|c) \leq 1.$$

Instead of relying solely on corpus-based estimates from a small parallel corpus, we employ a mixture model as follows:

$$P(e|c) = \beta P_{\text{corpus}}(e|c) + (1 - \beta)P_{\text{lexicon}}(e|c)$$

The retrieval results in Table 6 show that combining the probability estimates from the lexicon and the parallel corpus does improve retrieval performance. The best results are obtained when $\beta=0.7$; this is better than using uniform probabilities by 9% on Trec5C-medium and 4% on Trec6C-medium. Using the corpus probability estimates alone results in a significant drop in performance, the parallel corpus is not large enough nor diverse enough for reliable estimation of the translation probabilities. In fact, many words do not appear in the corpus at all. With a larger and better

parallel corpus, more weight should be given to the probability estimates from the corpus.

	Trec5-medium	Trec6-medium
P_lexicon	0.2449	0.3872
$\beta=0.3$	0.2557	0.3980
$\beta=0.5$	0.2605	0.4021
$\beta=0.7$	0.2658	0.4035
P_corpus	0.2293	0.2971

Table 6: Performance with different values of β . All scores are average precision.

11 Related Work

Other studies which view IR as a query generation process include Maron and Kuhns, 1960; Hiemstra and Kraaij, 1999; Ponte and Croft, 1998; Miller et al, 1999. Our work has focused on cross-lingual retrieval.

Many approaches to cross-lingual IR have been published. One common approach is using Machine Translation (MT) to translate the queries to the language of the documents or translate documents to the language of the queries (Gey et al, 1999; Oard, 1998). For most languages, there are no MT systems at all. Our focus is on languages where no MT exists, but a bilingual dictionary may exist or may be derived.

Another common approach is term translation, e.g., via a bilingual lexicon. (Davis and Ogden, 1997; Ballesteros and Croft, 1997; Hull and Grefenstette, 1996). While word sense disambiguation has been a central topic in previous studies for cross-lingual IR, our study suggests that using multiple weighted translations and compensating for the incompleteness of the lexicon may be more valuable. Other studies on the value of disambiguation for cross-lingual IR include Hiemstra and de Jong, 1999; Hull, 1997. Sanderson, 1994 studied the issue of disambiguation for mono-lingual IR.

The third approach to cross-lingual retrieval is to map queries and documents to some intermediate representation, e.g. latent semantic indexing (LSI) (Littman et al, 1998), or the General Vector space model (GVSM), (Carbonell et al, 1997). We believe our approach is computationally less costly than (LSI and GVSM) and assumes less resources (WordNet in Diekema et al., 1999).

12 Conclusions and Future Work

We proposed an approach to cross-lingual IR based on hidden Markov models, where the system estimates the probability that a query in one language could be generated from a document in another language. Experiments using the TREC5 and TREC6 Chinese test sets and the TREC4 Spanish test set show the following:

- Our retrieval model can reduce the performance degradation due to translation ambiguity. This had been a major limiting factor for other query-translation approaches.
- Some earlier studies suggested that query translation is not an effective approach to cross-lingual IR (Carbonell et al, 1997). However, our results suggest that query translation can be effective particularly if a bilingual dictionary is the primary bilingual resource available.
- Manual selection from the translations in the bilingual dictionary improves performance little over the HMM.
- We believe an algorithm cannot rule out a possible translation with absolute confidence; it is more effective to rely on probability estimation/re-estimation to differentiate likely translations and unlikely translations.
- Rather than translation ambiguity, a more serious limitation to effective cross-lingual IR is incompleteness of the bilingual lexicon used for query translation.

- Cross-lingual IR performance is typically 75% that of mono-lingual for our HMM on the Chinese and Spanish collections.

Future improvements in cross-lingual IR will come by attacking the incompleteness of bilingual dictionaries and by improved query expansion and context-dependent translation. Our current model assumes that query terms are generated one at a time. We would like to extend the model to allow phrase generation in the query generation process. We also wish to explore techniques to extend bilingual lexicons.

References

- L. Ballesteros and W.B. Croft 1997. "Phrasal translation and query expansion techniques for cross-language information retrieval." *Proceedings of the 20th ACM SIGIR International Conference on Research and Development in Information Retrieval* 1997, pp. 84-91.
- L. Ballesteros and W.B. Croft, 1998. "Resolving ambiguity for cross-language retrieval." *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998, pp. 64-71.
- J.P. Callan, W.B. Croft and J. Broglio. 1995. "TREC and TIPSTER Experiments with INQUERY". *Information Processing and Management*, pages 327-343, 1995.
- J. Carbonell, Y. Yang, R. Frederking, R. Brown, Y. Geng and D. Lee, 1997. "Translingual information retrieval: a comparative evaluation." *In Proceedings of the 15th International Joint Conference on Artificial Intelligence*, 1997.
- M. Davis and W. Ogden, 1997. "QUILT: Implementing a Large Scale Cross-language Text Retrieval System." *Proceedings of ACM SIGIR Conference*, 1997.
- A. Diekema, F. Oroumchain, P. Sheridan and E. Liddy, 1999. "TREC-7 Evaluation of Conceptual Interlingual Document Retrieval (CINDOR) in English and French." *TREC7 Proceedings*, NIST special publication.
- P. Fung and K. Mckeown. "Finding Terminology Translations from Non-parallel Corpora." *The 5th Annual Workshop on Very Large Corpora*, Hong Kong: August 1997, 192-202
- F. Gey, J. He and A. Chen, 1999. "Manual queries and Machine Translation in cross-language retrieval at TREC-7". *In TREC7 Proceedings*, NIST Special Publication, 1999.

- Harman, 1996. The TREC-4 Proceedings. NIST Special publication, 1996.
- D. Hiemstra and F. de Jong, 1999. "Disambiguation strategies for Cross-language Information Retrieval." *Proceedings of the third European Conference on Research and Advanced Technology for Digital Libraries*, pp. 274-293, 1999.
- D. Hiemstra and W. Kraaij, 1999. "Twenty-One at TREC-7: ad-hoc and cross-language track." In *TREC-7 Proceedings*, NIST Special Publication, 1999.
- D. Hull, 1993. "Using Statistical Testing in the Evaluation of Retrieval Experiments." *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 329-338, 1993.
- D. A. Hull and G. Grefenstette, 1996. "A dictionary-based approach to multilingual information retrieval". *Proceedings of ACM SIGIR Conference*, 1996.
- D. A. Hull, 1997. "Using structured queries for disambiguation in cross-language information retrieval." In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. AAAI, 1997.
- M. E. Maron and K. L. Kuhns, 1960. "On Relevance, Probabilistic Indexing and Information Retrieval." *Journal of the Association for Computing Machinery*, 1960, pp 216-244.
- D. Miller, T. Leek and R. Schwartz, 1999. "A Hidden Markov Model Information Retrieval System." *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214-221, 1999.
- D.W. Oard, 1998. "A comparative study of query and document translation for cross-language information retrieval." In *Proceedings of the Third Conference of the Association for Machine Translation in America (AMTA)*, 1998.
- Ari Pirkola, 1998. "The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval." *Proceedings of ACM SIGIR Conference*, 1998, pp 55-63.
- J. Ponte and W.B. Croft, 1998. "A Language Modeling Approach to Information Retrieval." *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275-281, 1998.
- L. Rabiner, 1989. "A tutorial on hidden Markov models and selected applications in speech recognition." *Proc. IEEE* 77, pp. 257-286, 1989.
- M. Sanderson. "Word sense disambiguation and information retrieval." *Proceedings of ACM SIGIR Conference*, 1994, pp 142-151.
- Voorhees and Harman, 1997. *TREC-5 Proceedings*. E. Voorhees and D. Harman, Editors. NIST special publication.
- Voorhees and Harman, 1998. *TREC-6 Proceedings*. E. Voorhees and D. Harman, Editors. NIST special publication.
- J. Xu and W.B. Croft, 1998. "Corpus-based stemming using co-occurrence of word variants". *ACM Transactions on Information Systems*, January 1998, vol 16, no. 1.