Structured Queries, Language Modeling, and Relevance Modeling in Cross-Language Information Retrieval

Leah S. Larkey and Margaret E. Connell
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
413-545-3415
Fax: 413-545-1789
{larkey|connell}@cs.umass.edu

Abstract

Two probabilistic approaches to cross-lingual retrieval are in wide use today, those based on probabilistic models of relevance, as exemplified by INQUERY, and those based on language modeling. INQUERY, as a query net model, allows the easy incorporation of query operators, including a synonym operator, which has proven to be extremely useful in cross-language information retrieval (CLIR), in an approach often called structured query translation. In contrast, language models incorporate translation probabilities into a unified framework. We compare the two approaches on Arabic and Spanish data sets, using two kinds of bilingual dictionaries – one derived from a conventional dictionary, and one derived from a parallel corpus. We find that structured query processing gives slightly better results when queries are not expanded. On the other hand, when queries are expanded, language modeling gives better results, but only when using a probabilistic dictionary derived from a parallel corpus.

We pursue two additional issues inherent in the comparison of structured query processing with language modeling. The first concerns query expansion, and the second is the role of translation probabilities. We compare conventional expansion techniques (pseudo-relevance feedback) with relevance modeling, a new IR approach which fits into the formal framework of language modeling. We find that relevance modeling and pseudo-relevance feedback achieve comparable levels of retrieval and that good translation probabilities confer a small but significant advantage.

Keywords: Crosslingual information retrieval, language modeling, structured query translation, query expansion.

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

1 Introduction

The central problem in information retrieval is ranking documents according to their relevance to a query. In cross-language retrieval, the documents are in one language and the queries are in another. Probabilistic models for information retrieval rank documents based on probabilities, or scores related to probabilities, in many different ways. Probabilistic models in wide use today fall into two major classes – traditional *retrieval models* which attempt to estimate probability of relevance of each document given a query, and *language models*, which attempt to model the generation of a query given a document.

Each of these approaches has its strengths and weaknesses. Both approaches can use multiple translations for query terms. Traditional retrieval models have been extended to handle multiple evidence and structured queries. Within systems based on these models, researchers have improved retrieval performance successfully by techniques such as query expansion and structured query translation, which are heuristic and ad hoc, rather than by extensions of the formal models. Language models provide a formal framework which gives more guidance for handling translation, query (and document) expansion, and other enhancements within the model. However, important elements like structured queries have not yet been incorporated.

One goal of this research is a careful comparison of these two different approaches to cross-language information retrieval. We compare two widely used approaches based on the two probabilistic retrieval models, the query net model with structured query translation, as implemented in INQUERY, and the crosslingual language model. The goal is not simply to make claims that one model or the other is better, since our experience shows that we can get comparable performance with both models. Rather, we explore the strength and weakness of each method, by examining how each performs with different kinds of resources. In addition, we compare two different approaches to query expansion and explore the role of translation probabilities in language modeling.

In what follows, we first review retrieval models and language modeling, and highlight the differences between these two approaches that are of interest in the present research. Next, we review two approaches to query expansion, one widely used and one new. Finally, we present the experiments on English/Arabic and English/Spanish cross-language retrieval and discuss the conclusions that can be drawn about the two approaches, about query expansion, and about translation probabilities.

1.1 Retrieval Models and Structured Query Translation

The first probabilistic retrieval model was published by Maron and Kuhns (1960). Their goal was to measure for each document the probability that the document will satisfy a given request for information. Other well known probabilistic models are those of Robertson and Sparck-Jones (1976), Croft and Harper (1979), Fuhr (1989) and Turtle and Croft (1991). These all share certain properties. The probability that a document is relevant to a query is a function of the distributions of query terms in relevant and non-relevant documents. The models differ in what is assumed about these distributions and how these probabilities are estimated.

INQUERY (Turtle and Croft, 1991), is unique in incorporating an inference net model, allowing structured queries. Like many other approaches, INQUERY computes a *belief* score that a

document D is relevant to a query Q, based on a weight for each query term t, typically a *tf·idf* score like this one in recent versions of INQUERY:

$$score_{t,D} = 0.4 + 0.6 \times \left(TF_{t,D} \times IDF_t \right) \tag{1}$$

$$TF_{t,D} = \frac{tf_{t,D}}{tf_{t,D} + 0.5 + 1.5 \frac{|D|}{\text{avg len}}} \tag{2}$$

$$IDF_t = \frac{\log \frac{N + 0.5}{|\{d_t\}|}}{\log N + 1} \tag{3}$$

where $tf_{t,D}$ is the number of occurrences of term t in document D, $|D|$ is the length of document D in words, N is the number of documents in the collection, $|\{d_t\}|$ is the size of the set of documents that contain term t, and *avg len* is the mean length of documents in the collection. The *TF* component of this weight is the Okapi *tf* (Robertson, et al., 1995). For a typical weighted sum query, the score for a document is simply a weighted average of term scores over the term occurrences in the query:

$$\text{Score}_{Q,D} = \frac{\sum_{t \in Q} w_t score_{t,D}}{|Q|} \tag{4}$$

However, the query net approach of Turtle and Croft (1991) allows the use of many operators. The #wsum operator computes the average shown in Equation (4). Other operators, such as Boolean AND and OR, and synonyms, combine term weights in more complex ways.

Of particular interest in the present work is the synonym operator. To treat a set of terms as synonyms, *tf·idf* weights are derived for the set of synonyms based on the statistics of the terms that make up the set, in effect treating the occurrences of all the synonyms in the set as if they were occurrences of a single word. Thus, if $T=\{t\}$ is a set of terms to be treated as synonyms,

$$tf_{T,D} = \sum_{t \in T} tf_{t,D} \tag{5}$$

$$\{d_T\} = \bigcup_{t \in T} \{d_t\} \tag{6}$$

where $tf_{T,D}$ is the number of occurrences of all of the terms in the set T in the document D, and $\{d_T\}$ is the set of documents containing any of the terms in T. This $tf_{T,D}$ and d_T can be substituted into Equations (2) and (3) to yield the *tf·idf* score for a synonym set.

Researchers have found the synonym operator useful for cross language retrieval. A successful approach has been to translate the query terms using a dictionary, and treat all the alternative translations for a word as a synonym set (Ballesteros and Croft, 1998; Pirkola, 1998). This

method is often called *structured query translation*. A simple example for Spanish can be seen in Figure 1.

```
#wsum (
 1 #syn ( polución )
 1 #syn ( azteca  charro guachinango mejicano mejicana mexicano mexicana pipil  )
 1 #syn  ( cabecer  cabeza  capi capital capitel chapitel corte mayúscula mayúsculo principal versal )
 )
```

Figure 1: Example of structured query translation for a query fragment *Pollution in the Mexican capital* using translations from a dictionary. Stop words have been removed.

1.2 Language Models

Language models (LM) have been used for a long time in speech recognition (Jelinek, Bahl, and Mercer, 1975; Bahl, Jelinek, and Mercer, 1983; Jelinek, 1997). More recently, they have made their appearance in the machine translation domain (Brown, et al., 1990; Brown et al., 1993), and then in information retrieval (Ponte and Croft, 1998; Berger and Lafferty, 1999; Song and Croft, 1999; Miller, et al, 1999).

A language model is a probability distribution over terms. In the simplest version of these models, the unigram model, the terms are single words or stemmed words. The particular form of language model we describe here is due to Miller, et al. (1999). A query is a bag or a sequence of single terms, generated from independent random samples of a term from one of two distributions – the distribution of words in a model of a document, and the distribution of words in a background model such as General English. For each pick, one samples the document model with probability λ, and the background model with probability 1-λ. Thus, the probability of generating a query Q from a document D and a background model GE is:

$$P(Q \mid D) = \prod_{e \in Q} \left(\lambda P(e \mid D) + (1 - \lambda) P(e \mid GE) \right) \tag{7}$$

where e is an English word in query Q, $P(e|D)$ is the probability of drawing word e from the document model, and $P(e|GE)$ is the probability of drawing word e from the background model of general English.

Several different crosslingual language models have been proposed. (Hiemstra and de Jong, 1999; Berger and Lafferty, 1999; Xu, et al., 2001; Federico and Bertoldi, 2002). Our crosslingual work is based on a widely-used extension of the above monolingual model (Xu et al., 2001). To generate a query from a document in a different language, say Arabic, one samples either the Arabic document, or the English background model. When the sample is taken from the Arabic document model, an Arabic word is first chosen at random from the document model, and then an English translation for that Arabic word is chosen at random from a bilingual lexicon. Thus the probability of generating an English query Q_e from an Arabic document model D_a is

$$P(Q_e \mid D_a) = \prod_{e \in Q_e} \left(\lambda \sum_{a \in Arabic} P(a \mid D_a) P(e \mid a) + (1 - \lambda) P(e \mid GE) \right) \tag{8}$$

where e, a, GE, $P(e|GE)$ and λ are as in Equation (7). $P(e|GE)$ is estimated as the relative frequency of English word e in some large sample of English documents. $P(a|D_a)$, the probability

of drawing Arabic word *a* from an Arabic document model D_a, is estimated as the relative frequency of Arabic word *a* in the document D_a. $P(e|a)$ is the conditional probability of choosing translation *e*, given Arabic word *a*. How $P(e|a)$ is estimated depends upon what kind of cross-language resources are available. If one has a sentence-aligned parallel corpus, one can estimate these probabilities by (appropriately smoothed) relative frequencies of alignments of words *e* and *a*. If a bilingual lexicon without alignment frequencies is all that is available, it is conventional to estimate $P(e|a)$ as $1/n$, where *n* is the number of English translations for the Arabic word *a*.

Researchers using language modeling have been able to attain performance comparable to that of *tf·idf* models like INQUERY, but without the extensive tuning of parameters. Language modeling has become popular because it is well grounded in statistical theory and is easily extendable to handle enhancements such as query and document expansion, n-gram units, stemming alternatives, etc.

1.3 Translation Probabilities

The *translation probability* $P(e|a)$ above is an important component of these language models. Good estimates of translation probabilities can be obtained by aligning parallel corpora, and counting the occurrences of alignments of word pairs. It is widely believed that the excellent cross-language results obtained in recent TREC experiments stem from the accurate estimation of translation probabilities obtainable from parallel corpora, and to the model's ability to use these translation probabilities. In contrast, structured query translation does not use translation probabilities. It is clear that some translations are more likely than others, and that structured query translation performance can be diminished by the inclusion of too many alternative translations. It is common practice, when starting with a dictionary made from a parallel corpus, to discard low probability translations.

If the translation probabilities are the determining factor in the effectiveness of language modeling, then retrieval should be less effective when only poor translation probability estimates are available. However, several researchers in our lab have mentioned paradoxical findings to us (Personal communication) in which large changes to translation probabilities made little difference in language model retrieval effectiveness. Furthermore, we have attained good results with conventional dictionaries in language modeling, using the relatively poor estimates of $1/n$ mentioned above. This paints a confusing picture, which we felt needed closer scrutiny.

1.4 Query Expansion and Relevance Modeling

An almost universal finding in IR, and CLIR, is that regardless of the model used, expansion techniques improve retrieval, as measured by an increase in mean average precision. Early work in informational retrieval showed that queries could be improved by *relevance feedback,* that is, adding and reweighting terms from retrieved documents relevant to a query (Roccio, 1971). Later work showed that the same general approach could be applied without relevance information, using a small number of the top-ranked retrieved documents (Xu and Croft, 1996). This *pseudo-relevance feedback* approach has been so successful that most of the best-performing systems at the TREC (Oard and Gey, 2003), CLEF (Peters, et al. 2002) and NTCIR (NTCIR Workshop, 2001) cross-language evaluations use it in some form.

Recently, researchers have proposed query expansion methods within the monolingual language modeling framework (Ponte, 1998; Zhai and Lafferty, 2001; Lavrenko and Croft, 2001). In cross-language retrieval, however, the researchers attaining the best crosslingual performance are still using the same local context analysis or pseudo-relevance feedback techniques used with

traditional retrieval models. These techniques do not fit into the formal framework of language modeling.

Lavrenko and Croft (2001) have recently developed a new kind of language model called a *relevance model* which brings the concept of relevance back into language modeling but which can also be viewed as a query expansion technique. In this framework we assume that for every information need there exists an underlying relevance model R which assigns the probabilities $P(w|R)$ – the probability of observing a word w in the set of documents relevant to that information need. The innovation in this approach is in assuming that both queries and relevant documents are random samples from the distribution $P(w|R)$. This is in contrast to other language model formulations which assume that only queries are generated by sampling document models, and to traditional retrieval models which estimate $P(w|R)$ heuristically. $P(w|R)$ is approximated by the probability of co-occurrence between the word w and the query $Q=e_1 \ldots e_k$.

$$P(w \mid R) = \frac{P(w, e_1 \ldots e_k)}{P(e_1 \ldots e_k)} \tag{9}$$

Lavrenko and Croft present two ways of estimating the joint probability $P(w, Q)$. In Method 1, used in the present research:

$$P(w, Q) = \sum_{D \in \mathrm{M}} P(D) P(w \mid D) \prod_{i=1}^{k} P(e_i \mid D) \tag{10}$$

where M is a set of unigram distributions we are sampling, $P(D)$ is the probability of choosing distribution D, $P(w|D)$ is the probability of choosing word w from the model D, and $P(e_i|D)$ is the probability of choosing query word e_i from model D. $P(D)$ is taken to be uniform, and both $P(w|D)$ and $P(e_i|D)$ are smoothed relative frequencies:

$$P(w \mid D) = \lambda \left(\frac{tf_{w,D}}{\sum_{v \in D} tf_{v,D}} \right) + (1 - \lambda) P(w \mid GE) \tag{11}$$

Here, λ is a smoothing parameter as in Equation (7), $P(w|GE)$ is the probability of word w in the background model as in Equation (7),and $tf_{w,D}$ is the frequency of word w in document D.

Note that this form of relevance modeling can be viewed as query expansion. The set of unigram distributions are document models, and in practice the set of documents is obtained by taking the top ranked documents from an LM retrieval pass.

Relevance models were extended to handle cross-language IR by Lavrenko, Choquette, and Croft (2002). Consider English queries and Arabic documents. The cross-language relevance model is given in Equation (12), which is the same as Equation (10), above, but now subscripts indicate that the documents D_a and terms w_a are in Arabic, and the query terms e_i are English:

$$P(w_a \mid R_a) = P(w_a, Q_e) = \sum_{D_a \in \mathrm{M}} P(D_a) P(w_a \mid D_a) \prod_{i=1}^{k} P(e_i \mid D_a) \tag{12}$$

$P(w_a|D_a)$ can still be estimated as in Equation (11), with the document and background models in Arabic. $P(e_i|D_a)$ is estimated as in Berger and Lafferty (1999) and Xu, et al., (2001):

$$P(e_i \mid D_a) = \lambda \sum_{a \in Arabic} P(a \mid D_a) P(e_i \mid a) + (1-\lambda) P(e_i \mid GE) \quad (13)$$

Note that this is the same as the expression inside the product in Equation (8). Once the relevance model is estimated, documents are ranked according to their cross-entropy with the relevance model:

$$CrossEntropy(R, D) = \sum_{w} P(w \mid R) \log P(w \mid D) \quad (14)$$

2 Overview of present and previous research

This research is intended to address the three issues discussed above. The first is a direct, well controlled comparison of the two dominant approaches to cross language retrieval, holding constant the kinds of preprocessing performed on the collections and queries.
In CLIR today, two of the probabilistic approaches reviewed above – structured query translation and language modeling are used widely (Oard and Gey, 2003). However, research groups tend to choose one approach or the other, and so direct comparisons have been rare.

One other study has compared structured query translation with LM for cross-language retrieval (Xu, et al, 2001). They found that language modeling performed better than structured query translation when used with a dictionary derived from a parallel corpus, or a combined dictionary derived from parallel and nonparallel sources. However, they also found that language modeling and structured query translation gave comparable results when used with dictionaries that did not have translation probabilities obtained from parallel corpora. But the story was not completely told. Their experiments covered only Chinese/English retrieval, and they presented results only with unexpanded queries. In order to make this comparison general, we use English queries with two very different languages, Arabic and Spanish. For each language, we use two different dictionaries, one probabilistic dictionary derived from a parallel corpus, and one conventional dictionary. We also test both unexpanded and expanded queries.

Second, we compare query expansion via pseudo-relevance feedback with relevance modeling. Lavrenko et al. (2002) compared relevance modeling with cross-lingual language modeling based on unexpanded queries, on the same Chinese data set used in the Xu, et al. (2001) experiments, but given that relevance modeling is effectively a query expansion technique, we feel that it is fairer to compare it with another expansion technique.

The third issue addressed by the present research is the importance of translation probabilities in the language modeling approach. The comparison of the two kinds of dictionaries has some bearing on this issue, but is confounded with differences in coverage. For this reason we include an experimental condition where probabilities from the parallel dictionary are discarded and replaced with probabilities of *1/n* for each of the *n* English translations for each Arabic or Spanish word in the parallel dictionary. If accurate translation probabilities are important, then this condition should show degraded performance relative to the condition using probabilities based on parallel data.

3 Experimental Methods

We present cross language retrieval experiments using two different retrieval systems, LM and INQUERY, performing identical tokenization, stemming, and stop word removal for both. The experiments are carried out with English queries and collections in two languages – Arabic and

Spanish. For each language, we have two sets of resources. The first is a probabilistic dictionary built from a parallel corpus, which has good translation probabilities. The second is a more conventional dictionary, which may or may not have good coverage, but does not have good probabilities. Monolingual retrieval conditions are included as baselines for crosslingual conditions.

3.1 Test Data

The corpus for the English-Arabic experiments, consisting of 383,872 Arabic documents from Agence France Presse from the years 1994-2000, was used for the TREC cross-language track in 2001 (Gey and Oard, 2002) and in 2002 (Oard and Gey, 2003). Title and description fields from two query sets were used - twenty-five queries from TREC 2001, and fifty queries from TREC 2002.

Two corpora and query sets were used for the English-Spanish experiments. From the TREC-4 multilingual track (Harman, 1996) we have El Norte newspaper articles from Mexico, and 25 TREC-4 topics provided both in Spanish and English. Queries for these experiments were title and description fields from the topics. The second English/Spanish data set was from TREC-5, for which the corpus was a set of Agence France Presse articles in Spanish from 1994 (Smeaton and Wilkinson, 1997). There were twenty-five topics (51-75) for this collection. Queries were made from the description fields because this set of topics had no Spanish titles.

A summary of the statistics of these data sets can be seen in Table 1. Note that the mean query lengths and document lengths do not include stop words.

Table 1 : Test Data Sets

	TREC 2001	**TREC 2002**	**TREC 4**	**TREC 5**
Corpus Language	Arabic	Arabic	Spanish	Spanish
Number of Documents	383,872	383,872	57,868	172,823
Size of Corpus (MB)	567	567	200	336
Mean doc len (words)	150	150	322	172
Query Languages	English, Arabic	English, Arabic	English, Spanish	English, Spanish
Number of queries	25	50	25	25
Mean query length in words	12.1 (English) 12.6 (Arabic)	11.0 (English) 11.2 (Arabic)	17.8 (English) 19.7 (Spanish)	7.2 (English) 7.7 (Spanish)
Relevant docs per query	165	118	160	100

3.2 Processing of Text

All of the English text (in queries, in parallel corpora, in dictionaries, and in the English collection used for English query expansion) was normalized to lower case but not stemmed. Stop words were removed, using INQUERY's stop list of 418 words. Numbers were also removed.

All the Arabic text was converted to Windows CP1256 encoding. Arabic text was then normalized and stemmed as described in Larkey, et al. (2003), removing punctuation, diacritics, non-letters, stop words, using a list of 168 Arabic stop words (Khoja and Garside, 1999), and

converting certain Arabic characters: replacing أ, إ, and آ with bare alif ا, replacing final ى with ي, and replacing final ة with ه. Then Arabic text was stemmed using the UMass light10 stemmer, which first strips و (*and*) from the beginnings of words, then removes definite articles (ال، وال، بال، كال، فال، لل) from word beginnings, and 10 suffixes from word ends (ها، ان، ات، ون، ين، يه، ية، ه، ة، ي) in a specific order allowing more than one suffix to be stripped as long as the suffix stripped later (more internal in the word) is lower on the list.

Spanish was normalized to lower case, stop words were removed, and words were stemmed using a Porter-like stemmer for Spanish (Broglio et al., 1995).

3.3 Bilingual Resources

The Arabic *parallel* or *probabilistic* dictionary was derived from the UN Arabic/English parallel corpus distributed by the Linguistic Data Consortium. It consists of 675 MB in 3,270,000 aligned sentences. English and Arabic sentences were processed as described above. The statistical translation training program GIZA++ (Och and Ney, 2000) was used to align the Arabic and English sentences, and to build a translation model using IBM model 4 (Brown et al, 1993). Translations which had a probability of less than .0001 were removed. A smaller version of each parallel dictionary was made for structured query translation, which contained only translation pairs with probabilities of .15 or higher, a choice based on our TREC2001 research. The sizes of parallel corpora and dictionaries are summarized in Table 2.

Table 2: Dictionaries used in Experiments

	Corpus		**Dictionary**		
Resource	**Size in MB**	**Number of parallel Sentences**	**Number of English Words**	**Number of Arabic or Spanish words**	**Number of translation pairs**
UN Parallel English/Arabic	675	3,270,000	114,534	156,290	1,832,268
Thresholded UN			71,305	145,038	310,286
Nonparallel Eng/Arabic			50,539	49,929	775,187
EP Parallel English/Spanish	240	746,000	63,750	37,868	841,255
Thresholded EP			32,413	32,469	71,778
Nonparallel Spanish/English			58,281	29,064	140,224

The *nonparallel* Arabic dictionary is the UMass dictionary, built for our TREC 2001 and 2002 work, from many different sources. It is described more thoroughly in Larkey, et al. (2003).

The Spanish parallel dictionary was built from a parallel corpus of European Parliament proceedings (Koehn, 2002) which are available in 11 languages. The Spanish and English part of the corpus consists of 240 MB of data in 746,000 aligned sentences, taken from proceedings between April, 1996 through December, 2001. We noticed some French files included both on the English and Spanish side, but we made no attempt to clean up the data. The Spanish parallel dictionary was built like the Arabic dictionary, using GIZA++.

The nonparallel Spanish dictionary was made from an electronic version of the Collins Spanish/English dictionary.

For both Arabic and Spanish, probability estimates of 1/*n* were assigned to translation pairs in the nonparallel UMass and Collins dictionaries, as described previously. For both Arabic and

Spanish, a composite dictionary was built by combining the parallel and nonparallel dictionaries. Each translation pair received the mean of its probability estimates from the component dictionaries

3.4 Retrieval Implementation

In this section we fill in implementation and parameter details. The experiments were carried out using our own search engine which can simulate INQUERY, and which can perform monolingual and crosslingual language modeling, all with the same preprocessing.

3.4.1 INQUERY and Structured Query Translation

For cross language retrieval, English queries were translated to Arabic or Spanish using structured query translation, according to the following procedure: For each English query word, if the word is found in the dictionary, take all the translations and place them inside a **#syn** (synonym) operator. If the English word is not found in the dictionary, stem the word with the *kstem* (Krovetz, 1993) stemmer, and try again. If any of the translations consist of a phrase rather than a single word, the phrase is enclosed in a **`#filreq`** operator. This operator is essentially a Boolean AND operator, which captures the requirement that if we are looking for "infantile paralysis," the document must contain both "infantile" and "paralysis." The final query is then a **#wsum** (weighted sum) of all the synonym sets as in Figure 1.

3.4.2 LM

LM retrieval was carried out using Equation (7) for monolingual, and Equation (8) for crosslingual retrieval. A value of λ=.5 was used in all monolingual runs, and λ=.7 was used for crosslingual runs. These values for λ, and the parameters for pseudo-relevance feedback (number of documents used for query expansion, number of words to add to expanded queries) were fixed at values that have worked well in past research. We did not tune these parameters for the present research. In crosslingual runs, if an English query word was not present in the dictionary, it was replaced with its stem and looked up again.

3.4.3 Query Expansion and Relevance Modeling

English queries were expanded using AP news articles from 1994 through 1998 from the Linguistic Data Consortium's North American News Supplement (LDC, 1998). Arabic and Spanish queries were expanded using the document collections being searched. To expand queries in INQUERY conditions, the top ranked 10 documents were taken from an INQUERY retrieval run, and all the words in the retrieved documents were ranked by the sum over the ten documents, of their *tf·idf* scores. For English queries (pre-translation), the top five new words were added to the queries. In expanding monolingual Arabic or Spanish queries, the top 50 new words were added to the queries. Final term weights were set to $2w_o+w_e$ where w_o is the original term weight, and w_e=1. Our pre-translation expansion is unusual in adding so few words. We have seen mixed success with pre-translation expansion when large numbers of word are added. Adding few words does not always aid retrieval as much as adding more words, but it rarely hurts performance.

In an expanded INQUERY cross-lingual run, the first pass was English query expansion. After the query was translated into a structured query with synonyms as in Figure 1, a post-translation expansion pass took the top 10 retrieved documents, and made a new structured query consisting of the old structured query in which each synonym set got twice its original weight, plus the 50

new terms were added under the weighted sum operator, each with a weight of 1. This expanded, translated, and further expanded query was used to retrieve the final ranked list of documents.

To expand queries in LM conditions with pseudo-relevance feedback, the top ranked 10 documents were taken from an LM retrieval run. The words from the retrieved documents were ranked, as in the INQUERY conditions, using *tf·idf* scores. For monolingual Arabic or Spanish, retrieval, the top 50 new terms were added to the original query, weighted $2w_o+w_e$ as above.

In an expanded LM crosslingual run using pseudo-relevance feedback, Arabic and Spanish query expansion was carried out by retrieving documents using the unexpanded English query. The top scoring 100 terms from the top ranked 10 documents were taken and used as a new Arabic or Spanish query (without the original terms), which was then run as a monolingual LM retrieval run. The final ranked list of documents for this LM run was combined with the ranked list from a cross-lingual run using the expanded English queries, and the final score for each document was the mean of its scores on the two ranked lists. Before combining the two ranked lists, scores were normalized according to the formula $score_{norm} = (score - min) / (max - min)$. Then scores were summed across the two lists.

In selecting values for the relevance modeling parameters, we were guided by experiments reported in Lavrenko, et al (2002) and Liu and Croft (2002). We did not tune the parameters on the present data. In a monolingual relevance model run, the first pass was an LM run with Dirichlet smoothing, $\lambda = doclength/(doclength+1000)$. For cross-lingual relevance modeling, the first pass is a crosslingual LM run. For both monolingual and crosslingual relevance modeling, a new "query" (relevance model) was made with 500 terms from the top 20 documents in the monolingual case, and the top 50 documents in the crosslingual case, weighted as described previously in Equation (12). 500 terms would be a large number for a query, but they allow a good estimate of the language model of relevant documents. Many of the added terms receive very low probabilities, so they do not have the detrimental effect on precision that the terms would have if added to a conventional query.

4 Comparison of Retrieval Approaches

The results of the experiments comparing retrieval approaches, and comparing expansion methods can be seen in Table 3-Table 6. Monolingual results are included as baselines. The Wilcoxon matched pairs test (Siegel, 1956) was used for all significance tests reported here. Results for the two Arabic query sets were combined and results for the two Spanish query sets were combined to give the statistical tests more power. A p-value of .05 was considered the cutoff for significance.

Table 3: English-Arabic CLIR. Mean average precision on 25 TREC 2001 queries

	Dictionary	Unexpanded		Expanded		
		INQ	LM	INQ	LM	Rel
Monolingual		.4135	.3792	.4367	.4174	.4109
Crosslingual	UMass (nonparallel)	.3807	.3132	.4518	.4047	.3897
	Parallel Corpus	.3160	.3057	.3812	.4131	.3958
	Combined	.3809	.3488	.4443	.4540	.4432
% of Mono	Combined	92	92	102	109	108

Table 4: English-Arabic CLIR. Mean average precision on 50 TREC 2002 queries

	Dictionary	Unexpanded		Expanded		
		INQ	LM	INQ	LM	Rel
Monolingual		.3225	.3084	.3623	.3708	.3598
Crosslingual	UMass (nonparallel)	.2947	.2688	.3330	.3446	.3473
	Parallel Corpus	.2872	.2940	.3413	.3538	.3637
	Combined	.3162	.3410	.3651	.3933	.3972
% of Mono	Combined	98	111	101	106	110

Table 5: English-Spanish CLIR. Mean average precision on 25 TREC 4 queries

	Dictionary	Unexpanded		Expanded		
		INQ	LM	INQ	LM	Rel
Monolingual		.4994	.4838	.5259	.5188	.4845
Crosslingual	Collins (nonparallel)	.3596	.3250	.3900	.4159	.4407
	Parallel Corpus	.4144	.4023	.4775	.4690	.4830
	Combined	.4024	.3972	.4464	.4681	.4813
%age of Mono	Parallel	83	83	91	90	100

Table 6: English-Spanish CLIR. Mean average precision on 25 TREC 5 queries

	Dictionary	Unexpanded		Expanded		
		INQ	LM	INQ	LM	Rel
Monolingual		.3906	.3793	.4714	.4543	.4778
Crosslingual	Collins (nonparallel)	.1912	.1722	.2700	.2667	.2652
	Parallel Corpus	.2815	.3707	.3535	.4419	.4871
	Combined	.2845	.3572	.3537	.4486	.4556
%age of Mono	Parallel	72	98	75	97	102

4.1 Monolingual results

Monolingual retrieval results on individual query sets can be seen in each of the tables above. On unexpanded queries, INQUERY performs significantly better than LM (Arabic $p=.0002$; Spanish $p<.05$). After query expansion, there are no significant differences among the monolingual conditions: Expanded INQUERY, LM with pseudo-relevance feedback, and relevance modeling all perform equivalently.

The monolingual retrieval results also confirm a pattern that is well known in the literature: query expansion generally improves retrieval performance independent of the retrieval model used.

4.2 Cross-language retrieval results

The cross-language retrieval results can be seen in Table 3-Table 6. The pattern of results on crosslingual retrieval is more complicated than on monolingual retrieval. For this reason, pairwise comparisons with significance levels are shown in Table 7. The parallel dictionaries show a different pattern of results than the probabilistic and combined dictionaries. The important points can be summarized as follows:

- Nonparallel dictionaries show a pattern of results similar to monolingual retrieval:
  - INQUERY (structured query translation) performs significantly better than LM on unexpanded queries.
  - On expanded queries, INQUERY with pseudo-relevance feedback, LM with pseudo-relevance feedback, and LM with relevance modeling, all perform equally well.
- Probabilistic and Combination Dictionaries
  - There is no consistent difference between INQUERY and LM on unexpanded queries.
  - LM with pseudo-relevance feedback and LM with relevance modeling are better than INQUERY with query expansion.
  - There is no consistent difference between the two LM expansion techniques. Relevance feedback appears to be significantly better on Spanish but not on Arabic.

Table 7: Significance tests on crosslingual comparisons. Asterisks indicate significant differences, question marks indicate a significant difference on Spanish data but not on Arabic.

Comparison		**p-values**	
	Dictionary	**Arabic**	**Spanish**
Inq > LM	nonprob	*< .01	*< .02
Inq-exp = LM-rel	nonprob	.50	.14
Inq-exp = LM-exp	nonprob	.29	.37
LM-exp = LM-rel	nonprob	.82	.06
Inq <? LM	prob	.48	*< .05
Inq-exp <? LM-rel	prob	.054	*< .01
Inq-exp <? LM-exp	prob	.08	*< .02
LM-exp <? LM_rel	prob	.71	*< .02
Inq = LM	combined	.39	.09
Inq-exp < LM-rel	combined	*<.02	*< .02
Inq-exp < LM-exp	combined	*< .05	*< .002
LM-exp = LM_rel	combined	.92	.16

We note other patterns in these data. Query expansion improves cross-lingual performance greatly. We also note that on Arabic, crosslingual retrieval is better than monolingual retrieval, even when we correctly use expanded monolingual as the baseline for expanded crosslingual, which many researches have not done when making such comparisons. Crosslingual performance as a percentage of monolingual can be seen in the last row of Tables 3-6. For Arabic these percentages refer to the combined dictionary. For Spanish the percentages refer to

the parallel dictionary. The nonparallel dictionary for Spanish performed so badly, particularly on TREC 5, that the performance on the combination dictionary was worse than on the parallel dictionary alone. This poor performance appears to be due primarily to the poor coverage of names in the Collins dictionary.

Getting better performance on crosslingual retrieval than on monolingual retrieval may seem paradoxical. A closer examination of one query sheds some light on how this can occur. TREC 2002 query 32 provides a good example. The English query is:

32. Caspian Beluga Conservation: What Beluga conservation projects are present in the Caspian region?

"Beluga" is obviously an important term in this query. In the Arabic query provided by NIST, "Beluga" is translated as بلوغا (blwga), essentially a transliteration of the word "Beluga." This Arabic word does not occur in the corpus. In the 34 documents that were judged relevant to this query, the Arabic word بلوغا does not occur at all. Instead, the articles use the words حفش (sturgeon) for Beluga, which is in the dictionary as a translation for sturgeon, and بيلوغا (bilwga), a slightly different transliteration than NIST's. Our English query expansion phase added the word "sturgeon" to the query, so the Arabic word حفش came into the translated expanded query. When we compare the monolingual performance to the crosslingual performance using the combined dictionary, we find .0715 monolingual average precision for query 32, and .8529 crosslingual LM average precision. Interestingly, in both monolingual and crosslingual cases, all 34 of the documents judged relevant were retrieved in the top 1000, but monolingual queries did a far poorer job at ranking them.

This example illustrates that dictionary translation can overcome a vocabulary mismatch problem. In general, using multiple translations is what allows dictionary-based CLIR methods to perform better than monolingual retrieval, particularly when coupled with expansion techniques that tend to emphasize words that co-occur with query terms. In monolingual search, a query may have apparently good search terms, but they may not happen to match the terms that occur in the corpus. This kind of mismatch is particularly likely with language pairs like English and Arabic, in which each has a high degree of variability in spelling foreign words.

In the Spanish experiments, and in Xu, et al.'s Chinese experiments, cross-language retrieval did not exceed monolingual. The difference may be due to the quality of the dictionaries, and whether NIST judged documents that were returned by queries in both languages, or just one language.

In this section we have presented the comparison of structured query translation and language modeling, and we have compared the two methods of query expansion. In the following section we focus on translation probabilities.

5 Role of translation probabilities

The crosslingual results above showed that when we have a parallel dictionary, language modeling can yield higher average precision in an IR task, presumably because LM is based on a model that incorporates translation probabilities, whereas INQUERY does not use translation probabilities. In this section we take a closer look at the importance of translation probabilities.

A superficial look at the results on the four data sets above shows that on the Spanish data sets, the parallel dictionary works far better than the conventional dictionary. On Arabic, in contrast,

there is little difference between the performance of the two dictionaries. However, the parallel vs non-parallel difference is confounded with differences in coverage. A more direct assessment of the importance of translation probabilities can be carried out by replacing the probabilities obtained from parallel corpus alignment with flat probability distributions.

The effect of probabilities is also confounded with the effect of thresholding. A dictionary made with a low threshold includes many low probability translation pairs. The language model uses the probabilities to limit the influence of these low probability translations on the final probability. A dictionary made with a higher threshold does not include these low probability translations. This leads to the question of whether including these words in the dictionary provides some benefit. In the INQUERY conditions, they have been thresholded out.

In the next experiment we examine LM performance using parallel dictionaries made with different thresholds, and we compare the results with performance on the dictionaries containing the same translation pairs, but with flattened probability distributions. For reference, we include INQUERY performance.

Four new dictionaries were made from each of the parallel dictionaries, by removing translation pairs where the probability $P(e|a)$ as in Equations (8) and (13) was less than .15, .1, .01, and .001. An additional set of dictionaries was made by replacing the probabilities in each of these dictionaries by $1/n$, where n is the number of translations of each word into English. The sizes of these dictionaries can be seen in Table 8 and Table 9.

Table 8: Sizes of thresholded dictionaries made from UN parallel English/Arabic corpus

Threshold	Number of English Words	Number of Arabic words	Number of translation pairs
.0001 (Full)	114,534	156,290	1,832,255
.001	109,821	156,290	1,503,999
.01	99,658	156,290	986,569
.1	79,319	153,694	418,445
.15	71,305	145,038	310,286

Table 9: Sizes of thresholded dictionaries made from the Europarl parallel English/Spanish Corpus

Threshold	Number of English Words	Number of Spanish words	Number of translation pairs
.0001 (Full)	63,589	38,121	841,255
.001	61,746	38,121	644,801
.01	54,724	38,121	339,311
.1	38,456	36,263	106,835
.15	32,413	32,469	71,778

The results of retrieval experiments using these dictionaries can be seen in Table 10. The thresholding results are similar to those found by Xu, et al (2001), in that LM degrades as the threshold is raised, and that INQUERY with structured query translation improves as the threshold is raised (within the range tested). When we compare the best performance of LM (at the lowest threshold) with the best performance of INQUERY (at high thresholds), the picture is

mixed. As we already noted in section 4.2, unexpanded queries show no consistent difference between LM and INQUERY. For expanded queries, LM is better than INQUERY.

Turning to the flat probabilities, the CLIR results can be seen in the columns labeled *LM 1/n* and *LM+rel 1/n.* In cells with an asterisk, the difference between full probabilities and flat probabilities is statistically significant. The pattern of results is consistent. With low thresholds, in which many low probability translations are included in the dictionary, higher average precision is obtained with full probabilities than with flat probabilities. At higher thresholds, the probabilities don't matter.

Overall, the most effect retrieval is attained using expanded queries with language modeling and translation probabilities derived by aligning a parallel corpus. One may note that the *tf·idf* and language models are conceptually similar. Both find a ranking score for a document in relation to a query, and that score is a function of relative term frequencies, smoothed by collection statistics.[1] However the crosslingual language model is more effective due to the added element of translation probabilities, $P(e|a)$. This suggests that one might be able to increase the effectiveness of structural query translation by incorporating translation probabilities into synonym processing.

Table 10: Mean Average Precision: Cross-language IR using parallel corpus dictionary comparing INQUERY, LM with actual probabilities and LM with probabilities flattened to *1/n*, across different thresholds. The boldface numbers show the results reported in Section 4.2. Asterisks indicate cases where there is a significant difference between the asterisked number and the precision in the cell immediately to the left.

		Unexpanded queries			Expanded Queries		
Data Set	**Threshold**	**INQ**	**LM**	**LM 1/n**	**INQ**	**LM+rel**	**LM+rel 1/n**
Arabic TREC2001	**.0001**		**.3057**			**.3958**	
	.001		.3025	.2806*		.3857	.3598*
	.01	.2312	.2878	.2761*	.3241	.3766	.3701*
	.1	.3106	.2478	.2499	.3789	.3283	.3268
	.15	**.3160**	.1806	.1810	**.3812**	.2657	.2655
Arabic TREC2002	**.0001**		**.2940**			**.3637**	
	.001	.1362	.2925	.2563*	.2566	.3639	.3430*
	.01	.2125	.2869	.2769*	.2990	.3664	.3549*
	.1	.2635	.2604	.2594	.3370	.3335	.3372
	.15	**.2872**	.2405	.2399	**.3413**	.3135	.3142
Spanish TREC4	**.0001**		**.4023**			**.4830**	
	.001	.3345	.3988	.3233*	.3912	.4816	.4259*
	.01	.3444	.3942	.3669*	.4506	.4808	.4558*
	.1	.4141	.3635	.3687	.4756	.4554	.4595
	.15	**.4144**	.3605	.3639	**.4775**	.4578	.4601
Spanish TREC5	**.0001**		**.3707**			**.4871**	
	.001	.2655	.3626	.2773*	.3388	.4841	.3560*
	.01	.2759	.3272	.2787*	.3220	.4288	.3861*
	.1	.3063	.2994	.2964	.3634	.4045	.4033
	.15	**.2815**	.2762	.2742	**.3535**	.3814	.3807

[1] For a more formal comparison, see Hiemstra and de Vries (2000).

6 Conclusions

In the comparison of structured query processing with language modeling, we found that structured query processing gave slightly better results than language modeling when queries were not expanded. On the other hand, when queries were expanded, language modeling gave better results, but only when using a probabilistic dictionary derived from a parallel corpus.

In comparing two methods of query expansion, we have shown that relevance modeling performed as well as pseudo-relevance feedback. On Spanish but not Arabic data, relevance modeling was significantly better than pseudo-relevance feedback. However, the processing of relevance modeling was substantially slower than pseudo-relevance feedback, so it is unclear whether it gives a practical advantage.

In examining the role of translation probabilities in a dictionary derived from a parallel corpus, we found that replacing the translation probabilities with flat probabilities results in a small but significant degradation in retrieval performance leading to the conclusion that accurate translation probabilities do contribute to higher precision.

7 References

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5*, 179-190.

Ballesteros, L., & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '98)* (pp. 64-71). Melbourne, Australia.

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd international conference on research and development in information retrieval* (pp. 222-229). Berkeley.

Broglio, J., Croft, W. B., Callan, J. P., & Nachbar, D. W. (1995). Document retrieval and routing using the INQUERY system. In D. K. Harman (Ed.), *Overview of the third Text REtrieval conference (TREC-3)* (pp. 29-38). Gaithersburg: National Institution of Standards and Technology Special Publication 500-225.

Brown, P. F., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., & Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics, 16*(2), 79-85.

Brown, P., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263-312.

Chen, A., & Gey, F. (2001). Translation term weighting and combining translation resources in cross-language retrieval, In E.M. Voorhees and D. K. Harman, (Eds.), *The tenth text*

retrieval conference, TREC 2001 (pp 529-533). Gaithersburg: National Institutes of Standards and Technology Special Publication 500-250.

.

Croft, W. B., & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation, 35*(4), 285-295.

Federico, M., & Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the twenty-fifth annual international ACM SIGIR conference on research and development in information retrieval* (pp. 167-174). Tampere, Finland.

Fuhr, N. (1989). Models for retrieval with probabilistic indexing. *Information Processing and Management, 25*(1), 55-72.

Gey, F. C., & Oard, D. W. (2002). The TREC-2001 Cross-language information retrieval track: Searching Arabic using English, French, or Arabic queries. In E.M. Voorhees and D. K. Harman, (Eds.), *The tenth text retrieval conference, TREC 2001* (pp 16-25). Gaithersburg: National Institutes of Standards and Technology Special Publication 500-250.

Harman, D. (1996). Overview of the fourth text retrieval conference (TREC-4). In D. Harman (Ed.), *The fourth text retrieval conference (TREC-4)*. Gaithersburg: National Institute of Standards and Technology Special Publication 500-236.

Hiemstra, D., & de Jong, F. (1999). Disambiguation strategies for cross-language information retrieval. In S. Abiteboul & A.-M. Vercoustre (Eds.), *Proceedings of the third European conference on research and advanced technology for digital libraries, ECDL '99.* (pp. 274-293). Paris, France.

Hiemstra, D., & de Vries, A. (2000). *Relating the new language models of information retrieval to the traditional retrieval models.* (CTIT Technical Report TR-CTIT-00-09). Enschede, The Netherlands: University of Twente. Available: http://wwwhome.cs.utwente.nl/~hiemstra/papers/

Khoja, S., & Garside, R. (1999). *Stemming Arabic text.* Lancaster, U.K.: Computing Department, Lancaster University. Available at http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, Massachusetts: MIT.

Jelinek, F., Bahl, L. R., & Mercer, R. L. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory, IT-21*, 250-256.

Koehn, P. (2002). *Europarl: A multilingual corpus for evaluation of machine translation*. Available: http://www.isi.edu/~koehn/publications/europarl/.

Krovetz, R. (1993). Viewing morphology as an inference process, In *Proceedings of the 16th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 191-203). Pittsburgh, PA.

Larkey, L. S., Allan, J., Connell, M. E., Bolivar, A., & Wade, C. (2003). UMass at TREC 2002: Cross language and novelty tracks. In *The eleventh text retrieval conference (TREC 2002)* (pp. 721-732) Gaithersburg: National Institute of Standards and Technology Special Publication 500-251.

Lavrenko, V., & Croft, W. B. (2001) Relevance-based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 120-127). New Orleans.

Lavrenko, V., Choquette, M., & Croft, W. B. (2002). Cross-lingual relevance models, In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 175-182). Tampere, Finland.

Liu, X., & Croft, W. B. (2002). Passage retrieval based on language models, In *Proceedings of the eleventh international conference on information and knowledge management, CIKM '02* (pp. 375-382). McLean, Virginia.

LDC (1998). Linguistic Data Consortium. North American News Text Supplement, LDC98T30. http://www.ldc.upenn.edu/Catalog/

Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery, 7*(3), 216-244.

Miller, D. R. H., Leek, T., & Schwartz, R. M. (1999). A hidden Markov model information retrieval system. In *Proceedings of the 22nd annual international conference on research and development in information retrieval* (pp. 214-221). Berkeley, California.

NTCIR Workshop 2. (2001) Proceedings of the second NTCIR workshop on research in Chinese and Japanese text retrieval and text summarization. Tokyo: National Institute of Informatics. Available: http:/research.nii.ac.jp/ntcir/workshop/OnlineProceedings2.

Oard, D. W., & Gey, F. C. (2003). The TREC-2002 Arabic/English CLIR track. In *The eleventh text retrieval conference (TREC 2002)*(pp. 17-26). Gaithersburg: National Institute of Standards and Technology Special Publication 500-251.

Och, F. J., & Ney, H. (2000). Improved statistical alignment models, In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics* (pp. 440-447). Hongkong, China.

Peters, C. (2001). *Cross-Language information retrieval and evaluation, Workshop of Cross-Language Evaluation Forum, CLEF 2000,* Lisbon, Portugal. Springer Verlag.

Peters, C., Braschler, M., Gonzalo, J., & Kluck, M. (2002). *Evaluation of cross-language information retrieval systems: Second workshop of the cross-language evaluation forum, CLEF 2001.* Darmstadt, Germany. Springer Verlag.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval, In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 55-63). Melbourne, Australia.

Ponte, J. M. (1998). *A language modeling approach to information retrieval.* Unpublished doctoral dissertation, University of Massachusetts, Amherst, MA.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval, In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 275-281). Melbourne, Australia.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129-146.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. In D. K. Harmon (Ed.), *Overview of the third text retrieval conference (TREC-3)* (pp. 109-126) Gaithersburg, MD: NIST special publication 500-225.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system - Experiments in automatic document processing*. Englewood Cliffs: Prentice Hall.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Smeaton, A., & Wilkinson, R. (1997). Spanish and Chinese document retrieval in TREC-5. In E. M. Voorhees & D. K. Harman (Eds.), *The fifth text retrieval conference (TREC-5)* (pp. 57-64). Gaithersburg: NIST Special Publication 500-238.

Song, F., & Croft, W. B. (1999). A general language model for information retrieval, *Proceedings of the eighth international conference on information and knowledge management, CIKM '99* (pp. 316-321). Kansas City, Missouri.

Turtle, H., & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems, 9*(3), 187-22.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis, In *Proceedings of the 19th international ACM SIGIR conference on research and development in information retrieval* (pp. 4-11). Zurich, Switzerland: ACM.

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval, In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval.* (pp. 105-110). New Orleans: ACM Press.

Zhai, C., & Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval, In *Proceedings of the tenth international conference on information and knowledge management, CIKM '01* (pp. 403-410). Atlanta.