Cross-Lingual Relevance Models

Victor Lavrenko, Martin Choquette and W. Bruce Croft
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003
{lavrenko,choquett,croft}@cs.umass.edu

ABSTRACT

We propose a formal model of Cross-Language Information Retrieval that does not rely on either query translation or document translation. Our approach leverages recent advances in language modeling to directly estimate an accurate topic model in the target language, starting with a query in the source language. The model integrates popular techniques of disambiguation and query expansion in a unified formal framework. We describe how the topic model can be estimated with either a parallel corpus or a dictionary. We test the framework by constructing Chinese topic models from English queries and using them in the CLIR task of TREC9. The model achieves performance around 95% of the strong mono-lingual baseline in terms of average precision. In initial precision, our model outperforms the monolingual baseline by 20%. The main contribution of this work is the unified formal model which integrates techniques that are essential for effective Cross-Language Retrieval.

Categories and Subject Descriptors

H.3.3 [**Information storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Experimentation

Keywords

Cross-Language Information Retrieval, Language Models

1. INTRODUCTION

The task of Cross-Language Information Retrieval (CLIR) addresses a situation when a query is posed in one language but the system is expected to return the documents written in another language. The scenario could have been considered unlikely a mere decade ago, but the explosive growth of the World Wide Web has blurred national boundaries to the point where a casual user may find an interest in retrieving documents in a foreign language. Once a user obtains a set of relevant documents in a foreign language, she can use automatic machine translation software to get a sense of the content. What remains is the problem of retrieving that set of documents, starting with a query in the user's native language.

In recent years, the problem of Cross-Language Retrieval has enjoyed significant interest from the research community, and a number of techniques were proposed to solve the problem [2, 9, 19]. Most of these techniques center around a common idea: they attempt to *translate* the query from the user's language to the language of the documents. In most cases, the translation is done in a word-by-word fashion, using a dictionary, a machine translation system, or a similar resource. Typically, any given word may have multiple possible translations, so significant effort has been devoted to disambiguating the resulting translations, either through the use of context [19], statistical co-occurrence [3, 19], triangulated translation [8], or a number of similar techniques. In addition, researchers found that in many cases it is helpful to include words that are not direct translations of any query word, but are closely related to the meaning of the query. This observation led to the common use of heuristic *query expansion* techniques [2, 19].

In this paper, we propose a model of Cross-Language Retrieval that does not rely on a word-by-word translation of the query. Instead, we attempt to construct an accurate relevance model in the target language, and use that model to rank the documents in the collection. Following the work of Lavrenko and Croft [12], we use the term **relevance model** to refer to a probability distribution, which specifies how often we expect to see any given word in the documents relevant to the query. To estimate relevance models in a cross-lingual setting we extend the methods proposed by [12], and show how the estimation can be done with either a parallel corpus or a dictionary. The main contribution of this work is the unified formal framework for integrating techniques that are essential for effective Cross-Language Retrieval, such as query expansion for dealing with synonymy, and translation disambiguation for handling polysemy.

The remainder of the paper is structured as follows. Section 2 briefly surveys recent developments in the fields of Cross-Language Retrieval and statistical Language Modeling. Section 3 describes how relevance models can be estimated in a cross-lingual environment, and how we can use them for retrieving documents. Section 4 addresses specific

SIGIR'02, August 11-15, 2002, Tampere, Finland

implementation issues that arose in working with Chinese. In Section 5 we test the performance of our model on the cross-language retrieval task of TREC9, and compare our performance with results reported by other researchers.

2. RELATED WORK

Previous applications of language modeling to cross-language retrieval have been reported by Hiemstra and de Jong [9] and Xu *et al.* [18, 19]. Although the model proposed by Berger and Lafferty [4] applies to the "translation" of a document into a query in a monolingual environment, it can readily accommodate a bilingual environment. The three approaches above all make use of translation probabilities attached to pairs of words. The pairs of words and their corresponding probabilities are often obtained from a bilingual dictionary by assigning the same probability to all the translations of a word. When a parallel corpus or a pseudo-parallel corpus (where parallel documents are produced by an MT system) is available, the required translations and probabilities can be obtained by applying Brown *et al.*'s approach to machine translation [5] (or some related technique [10]). Xu *et al.* [18] showed that combining statistics from various lexical resources can help correct problems of coverage and lead to significant improvements.

When no parallel corpus is available, our model uses the same estimation techniques as the other language modeling approaches. Unlike those approaches, however, our model does not rely on word-by-word translation when a comparable corpus or, *a fortiori*, a parallel corpus is available. [1] In such a case, our approach is closer in spirit to that of Sheridan and Ballerini [15] where the retrieval process consists of two retrieval steps. First, the best matching documents in the source language are retrieved. A query is then built by selecting words occurring frequently in the documents in the target language that are comparable to the ones retrieved. Finally, the new query is used to retrieve documents in the target language. Our approach does not select new query terms though. Instead, it uses the set of comparable documents to estimate, for each word in the target vocabulary, the probability of observing the word in the set of relevant documents.

The use of additional disambiguation techniques such as those based on phrases and co-occurrence measures (e.g., mutual information) still have to be explored within our framework. These techniques have been put to good use in approaches that do not rely on language models (e.g., [3, 7]).

3. MATHEMATICAL FRAMEWORK

Let $Q = e_1 \dots e_k$ be the query in the source language and let R_Q be the set of target documents that are relevant to that query. Lavrenko and Croft [12] suggest that effective ranking of target documents could be achieved if we had a way of estimating the *relevance model* of Q, i.e. the set of probabilities $P(w|R_Q)$ for every word w in the target vocabulary. $P(w|R_Q)$ denotes the probability that a word sampled at random from a relevant document would be the word w. If we knew what documents comprised R_Q, estimation of these probabilities would be straightforward, but in a typical retrieval environment we are not given any examples of relevant documents. Lavrenko and Croft [12] argue that in the absence of training data, a reasonable way to approximate $P(w|R_Q)$ is by a joint probability of observing the word w together with query words $e_1 \dots e_k$:

$$P(w|R_Q) \approx P(w|Q) = \frac{P(w, e_1 \dots e_k)}{P(e_1 \dots e_k)} \quad (1)$$

3.1 Mono-lingual estimation

Lavrenko and Croft [12] describe two methods for estimating the joint probability $P(w, e_1 \dots e_k)$ for the mono-lingual case, where w and $e_1 \dots e_k$ are words from the same vocabulary. Both methods assume there exists a set $\mathcal{M}$ of underlying source distributions from which w and $e_1 \dots e_k$ could have been sampled. The two methods differ in the kinds of independence assumptions they make. In this paper we will consider only *Method 1* because of its relative simplicity, and its decomposability, which will be leveraged in section 4.1. *Method 1* assumes w and $e_1 \dots e_k$ to be mutually independent once we pick a source distribution from $\mathcal{M}$, leading to the following estimate:

$$P(w, e_1 \dots e_k) = \sum_{M \in \mathcal{M}} P(M) \left(P(w|M) \prod_{i=1}^{k} P(e_i|M) \right) \quad (2)$$

Here $P(M)$ denotes some prior distribution over the set $\mathcal{M}$ (usually taken to be uniform), while $P(w|M)$ specifies the probability of observing w if we sampled a random word from M (and similarly for $P(e_i|M)$). We can take the universe $\mathcal{M}$ to be the set of documents in the database, and assume that w and all e_i are identically distributed with probabilities estimated using a simple smoothing of the relative frequency:

$$P(w|M_D) = \lambda \left(\frac{tf_{w,D}}{\sum_v tf_{v,D}} \right) + (1 - \lambda) P(w) \quad (3)$$

Here λ is a tunable parameter which determines the degree of smoothing. $tf_{w,D}$ is the number of times the word w occurs in document D. $P(w)$ is the background probability of observing the word w, obtained from a large corpus. Equation (3) is used to estimate $P(e_i|M_D)$ in the same way as $P(w|M_D)$.

3.2 Cross-lingual estimation

In the remainder of this section we describe how to extend the estimation to the cross-lingual case, when, for example, w is a Chinese word, and $e_1 \dots e_k$ are English words. In this case we obviously cannot assume that w and e_i are identically distributed. We discuss two estimation strategies, one based on a parallel corpus of documents, and one based on a bilingual lexicon.

3.2.1 Estimation with a parallel corpus

Suppose we have at our disposal a parallel corpus, a set of document pairs $\{E, C\}$, where E is an English document, and C is a Chinese document discussing the same topic. We let $\mathcal{M}$ be the set of corresponding distribution pairs

[1] Generally, comparable corpora are sets of topically related documents written in different languages. We use a more restricted definition, which mandates links between topically-related documents. Linked documents may not be exact translations of each other.

$\{M_E, M_C\}$, and estimate the joint probability of observing together w and $e_1 \dots e_k$ as:

$$P(w, e_1 \dots e_k) = \sum_{\{M_E, M_C\} \in \mathcal{M}} P(\{M_E, M_C\}) \left(P(w|M_C) \prod_{i=1}^{k} P(e_i|M_E) \right) \quad (4)$$

Here $P(\{M_E, M_C\})$ can be kept uniform, and $P(w|M_C)$ and $P(e_i|M_E)$ are calculated according to equation (3), but each in its own language.

3.2.2 Estimation with a bilingual lexicon

If no parallel corpus is available, it is still possible to estimate the joint probability $P(w, e_1 \dots e_k)$ if we have a statistical lexicon, which gives the translation probability $P(e_i|w)$ for every English word e_i and every Chinese word w. Note that any bilingual dictionary can be turned into a statistical lexicon by simply assigning uniform translation probabilities to all English translations of a given Chinese word. While the quality of this lexicon may be inferior, it will still serve the purpose. In this case we let $\mathcal{M}$ be the set of documents in Chinese. $P(w|M_C)$ can be computed directly from equation (3). In order to compute $P(e_i|M_C)$ for an English word e_i in a Chinese document C, we can use the translation model, advocated by Berger and Lafferty [4] and recently used in cross-lingual setting by [9, 19]. According to [19]:

$$P(e_i|M_C) = (1 - \lambda) P(e_i) + \lambda \sum_v P(e_i|v) P_{ml}(v|M_C) \quad (5)$$

Here the summation goes over all the Chinese words v in the vocabulary, $P(e_i|v)$ is the translation probability from the statistical lexicon, and $P_{ml}(v|M_C)$ is simply the number of times v occurs in C, divided by the length of C. $P(e_i)$ is the background probability of e_i, computed over a large corpus.

3.3 Ranking with a Relevance Model

In the previous section we proposed a technique for estimating a Chinese relevance model, starting with an English query. What remains is to specify a document ranking method, which will allow us to use the constructed relevance model to retrieve Chinese documents. Lavrenko and Croft [12] advocate using the Probability Ranking Principle [14], where documents are ranked by the probability ratio: $P(D|R)/P(D|N)$. In our experiments we found that a more stable metric is the relative entropy (also known as Kullback-Leibler divergence), suggested by Lafferty and Zhai in [11]. They formulate the retrieval problem as that of a risk minimization, and provide justification for using relative entropy as a risk metric between two distributions. Lafferty and Zhai [11] used KL divergence with their own technique for estimating query (relevance) models, which is Markov chains on inverted indices. We found that relative entropy works well with estimation suggested by Lavrenko and Croft [12], as well as with our extension of that work. The relative entropy between a relevance model R and a document model D is defined as:

$$KL(R||D) = \sum_w P(w|R) \log \frac{P(w|R)}{P(w|D)} \quad (6)$$

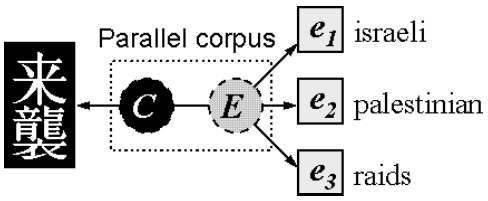

Figure 1: Graphical representation of cross-lingual relevance models. The model consists of a set of paired Chinese / English distributions. English queries are random samples from some English distribution from the paired set. Relevant Chinese documents are random samples from the corresponding Chinese distribution. Paired distributions could be estimated from a parallel corpus.

Documents are ranked in the order of increasing divergence, i.e. documents that have a smaller divergence with the Relevance Model are considered more relevant. In equation (6) $P(w|R)$ is computed from equation (1), as described above. $P(w|D)$ is calculated according to equation 3.

3.4 A Brief Summary of the Model

In this section we outlined a formal method for constructing a topic model in the target language, starting with a query in the source language. We also suggested using KL divergence with the topic model as a document ranking function. Massive query expansion is an integral part of the technique, since we compute the probability $P(w|R)$ for *every* word in the target language. Translation disambiguation is achieved automatically, because we compute co-occurrence (joint probability) between any word w and all the query words. Figure 1 shows a graphical representation of the underlying generative model.

4. IMPLEMENTATION

In the previous section we introduced cross-lingual relevance models, described how they can be estimated using either a parallel corpus or a bilingual lexicon, and described how the models could be used for retrieval purposes. In this section we describe how the system was implemented in practice, as well as specific processing steps that were performed on each language, and the resources that were used.

4.1 Making the Model Tractable

As specified, our model of estimation involves computing equation (2) or equation (4) over every document (or document pair) in the dataset. That computation has to be repeated for every word w in the Chinese vocabulary, which makes the estimation extremely expensive. In reality, it is possible to re-write equations (1) and (2) in a form that makes estimation feasible. If we expand the probability of the query as $P(e_1 \dots e_k) = \sum_w P(w, e_1 \dots e_k)$, and substitute equation (2) for the joint probability $P(w, e_1 \dots e_k)$, it becomes possible to estimate the probability of observing w in the set of relevant documents R_Q as:

$$P(w|R_Q) = \sum_{M \in \mathcal{M}} P(w|M) P(M|e_1 \dots e_k) \quad (7)$$

	LDC	CETA	HK News	Combined
English terms	86,000	35,000	21,000	104,997
Chinese terms	137,000	202,000	75,000	305,103

Table 1: Composition of the BBN bilingual lexicon

	HK News	TDT	HK News + TDT
Document pairs	18,147	46,692	64,839
English terms	28,806	67,542	83,152
Chinese terms	49,218	111,547	132,453

Table 2: Composition of the parallel corpus used in our experiments.

When expressed like this, it becomes obvious that a relevance model (under estimation method 1) is a linear mixture of distributions from $\mathcal{M}$, where each distribution M is "weighted" by its posterior probability of generating the query: $P(M|e_1 \dots e_k)$. In our model, the posterior probability is expressed as:

$$P(M|e_1 \dots e_k) = \frac{P(M) \prod_i P(e_i|M)}{\sum_M P(M) \prod_i P(e_i|M)} \qquad (8)$$

In practice, because of the product in the numerator, this posterior has near-zero values for all but a few models M in a given collection $\mathcal{M}$. These models are precisely the models that rank highest when the query $e_1 \dots e_k$ is issued against the collection $\mathcal{M}$. Therefore, instead of computing equation (2) over the entire collection, we can compute equation (7) over some number n of top-ranked models retrieved by $e_1 \dots e_k$.

Expressing the relevance model in terms of equation (7) has another advantage. We could relax the strict probabilistic interpretation of the posterior $P(M|e_1 \dots e_k)$ and substitute any heuristic estimate, as long as it is non-negative and sums to 1. In this way, relevance models could be constructed from any ranked list of documents.

4.2 Resources

All of our experiments were performed on the dataset used in the TREC9 cross-lingual evaluation. The dataset consists of 127,938 Chinese documents, totaling around 100 million characters. We used the official set of 25 queries. We used two query representations: *short* queries used only the title field, while *long* queries used title, description and narrative fields.

Experiments involving a bilingual dictionary used the statistical lexicon created by Xu et.al [19]. The lexicon was assembled from three parts: the LDC dictionary, the CETA dictionary, and the statistical dictionary, learned from the Hong-Kong News corpus by applying the GIZA machine translation toolkit. Table 1 provides a summary of the dictionary components.

In the experiments that made use of the parallel corpus, we used the Hong-Kong News parallel dataset, which contains 18,147 news stories in English and Chinese. Because it is so small, the Hong-Kong parallel corpus has a significant word coverage problem. In order to alleviate the problem, we augmented the corpus with the TDT2 and TDT3 [**?**] pseudo-parallel datasets. These corpora contain 46,692 Chinese news stories along with their SYSTRAN translations into English. Since the documents are translated by software, we do not expect the quality of the TDT corpus to be as high as Hong-Kong News. We discuss the impact of adding the TDT corpus in section 5. The composition of the parallel corpus is detailed in Table 2.

4.3 English processing

The English portions of the dataset were pre-processed as follows. Both the documents and the queries were tokenized on whitespace and punctuation. Tokens with fewer than two characters were discarded. A total of 400 stopwords from the InQuery [1] stoplist were removed. We used the *kstem* stemmer, developed by Krovetz, to normalize the word forms in all documents and queries. As an exception, we used the *Porter* stemmer [13] on experiments that used the bilingual dictionary, due to the fact that the BBN statistical lexicon [19] was Porter-stemmed. No other form of processing was used on either the queries or the documents.

4.4 Chinese processing

The pre-processing performed on the Chinese part of the corpus was very crude, due to our limited knowledge of the language. The entire dataset, along with the Chinese queries was converted into the simplified encoding (GB). We carried out separate experiments with three forms of tokenization: (i) single Chinese characters (unigrams), (ii) half-overlapping adjacent pairs of Chinese characters (bigrams), and (iii) Chinese "words", obtained by running a simple dictionary-based segmenter, developed by F. F. Feng at the University of Massachusetts. In section 5 we report separate figures for all three forms of tokenization, as well as a linear combination of them. We did not remove any stopwords, or any punctuation characters from either Chinese documents or queries. This results in some spurious matches and also in these characters figuring prominently in the relevance models we constructed.

4.5 Example

Figure 2 provides an example of the relevance model estimated from query number 58: *"environmental protection laws"*. We show 20 tokens with highest probability under the model. It is evident that many stopwords and punctuation characters are assigned high probabilities. This is not surprising, since these characters were not removed during pre-processing, and we naturally expect these characters to occur frequently in the documents that discuss any topic. However, the model also assigns high probabilities to words that one would consider highly relevant to the topic of environmental protection.

5. EXPERIMENTAL RESULTS

In this section we carry out an evaluation of the proposed model of cross-language retrieval on the TREC9 dataset. We compare the performance of the following models:

1. **Mono-lingual baseline.** We use the basic language modeling system, which was reported as a baseline in a number of recent publications [19, 17]. The Chinese documents D are ranked according to the probability that a Chinese query $c_1 \dots c_k$ was generated from the document model M_D. Word probabilities are estimated according to equation (3).

2. **Mono-lingual Relevance Model.** This system is included as an alternative mono-lingual baseline, and

Q = "environmental protection laws" 环境保护法

P(word|Q)	word	meaning
0.061	,	[punctuation]
0.036	的	[possessive suffix]
0.027	。	[punctuation]
0.017	和	and
0.016	、	[punctuation]
0.009	环境	environment
0.009	了	[end of sentence]
0.008	海洋	sea
0.008	法	law
0.008	资源	resource
0.007	全国	whole country
0.007	在	in
0.006	保护	protect
0.006	污染	pollution
0.006	胶	rubber
0.006	发泡	defects in plastic
0.005	与	and
0.005	中国	china
0.005	产品	product
0.005	法律	law

Figure 2: Example of a cross-lingual relevance model, estimated from query number 58. Shown are the 20 tokens with highest probabilities under the model.

to demonstrate the degree to which Relevance Models degrade, when estimated in a cross-lingual setting. Given a Chinese query $c_1 \dots c_k$, we estimate a Relevance Model as suggested by Lavrenko and Croft [12]. We used estimation method 1 (equation 2). We used relative entropy (equation 6) as the document ranking function.

3. **Probabilistic Translation Model.** As a cross-lingual baseline, we report the performance of our implementation of the system used by Xu et.al. [19]. The translation model was originally proposed by Berger and Lafferty [4] and Hiemstra and de Jong [9]. We used the formulation advocated by Xu et al. [19]. We used the same statistical lexicon and the same system parameters that were reported in [19].

4. **Cross-lingual Relevance Model (parallel).** Given an English query $e_1 \dots e_k$, we estimate a Relevance Model in Chinese using equations (1) and (4). We use the combined parallel corpus for estimating equation (4). The Chinese documents are then ranked by their Kullback-Leibler divergence from the Relevance model (equation 6).

5. **Cross-lingual Relevance Model (dictionary).** We estimate the cross-lingual relevance model using equations (1) and (2). Equation (5) is used to compute the probability of an English word e_i given a Chinese document C. We use the lexicon reported in [19] for translation probabilities $P(e_i|v)$. The documents are ranked by KL divergence from the Relevance Model (equation 6).

In all cases we performed separate experiments on the three representations of Chinese: unigrams, bigrams and "words". Refer to section 4.4 for details. The smoothing parameter λ was tuned separately for each representation as we found that smoothing affects unigrams and bigrams very differently. The results from the three representations were then linearly combined. The weights attached to each representation were set separately for every model, in order to show best results. As an exception, the Probabilistic Translation Model was evaluated on the same representation that was used by Xu et.al.[19]. Due to the absence of the training corpus, the tuning of all parameters was performed on the testing data using a brute-force hill-climbing approach. The small number of queries in the testing dataset precluded the use of any statistical significance tests.

5.1 Baseline Results

Table 3 shows the retrieval performance, of the described models on the TREC9 cross-language retrieval task. We use non-interpolated average precision as a performance measure. Percentage numbers indicate the difference from the mono-lingual baseline. We show results for both short and long versions of the queries. Our monolingual results form a strong baseline, competitive with the results reported by [17, 19]. This is somewhat surprising, since our processing of Chinese queries was very simplistic, and a lot of spurious matches were caused by punctuation and stopwords in the queries. We attribute the strong performance to the careful selection of smoothing parameters and combination of multiple representations.

Monolingual Relevance Model provides an even higher baseline for both short and long queries. The difference is highlighted in Figure 3. Relevance Models show good performance at higher recall, which is expected, since they can be interpreted as theoretically grounded query expansion techniques.

The Probabilistic Translation Model achieves around 85% - 90% percent of the mono-lingual baseline. Xu et.al. in [19] report the performance of the same model to be somewhat higher than our implementation (0.3100 for long queries). We attribute the differences to the different form of preprocessing used by Xu et.al, since we used the same bilingual lexicon and the same model parameters as [19].

5.2 Cross-lingual Relevance Model Results

Table 3 shows that Cross-lingual Relevance Models perform very well, achieving 93% - 98% of the mono-lingual baseline on the combined representation. This performance is better than most previously-reported results [17, 19], which is somewhat surprising, given our poor pre-processing of Chinese. Our model noticeably outperforms the Probabilistic Translation Model on both long and short queries (see figure 4). It is also encouraging to see that Cross-lingual Relevance Models perform very well on different representations of Chinese, even though they do not gain as much from the combination as the baselines.

Note that Relevance Models estimated using a bilingual lexicon perform better than the models estimated from the parallel corpus. We believe this is due to the fact that our parallel corpus has an acute coverage problem. The bilingual dictionary we used [19] covers a significantly larger number of both English and Chinese words. In addition, two thirds of our parallel corpus was obtained using automatic machine translation software, which uses a limited vocabulary. It is also worth noting that the remaining part of our parallel corpus, Hong-Kong News, was also used by Xu et al. [19] in the construction of their bilingual dictionary.

Average Precision	HK News	HK News + TDT	
Unigrams	0.1070	0.2258	+111%
Bigrams	0.1130	0.2519	+123%
"Words"	0.1210	0.2493	+106%

Table 5: Parallel corpus size has a very significant effect on the quality of Cross-lingual Relevance Models.

Table 5 illustrates just how serious the coverage problem is. We show performance of Relevance Models estimated using just the Hong-Kong News portion of the corpus, versus performance with the full corpus. We observe tremendous improvements of 100% to 200% percent, by adding the TDT data, even though this data was automatically generated using SYSTRAN.

5.2.1 High-precision performance

Average precision is one of the most frequently reported metrics in cross-language retrieval. This metric is excellent for research purposes, but it is also important to consider user-oriented metrics. Table 4 shows precision at different ranks in the ranked list of documents. Precision at 5 or 10 documents is what affects a typical user in the web-search setting. We observe that Cross-lingual Relevance Models exhibit exceptionally good performance in this high-precision area. Models estimated using the parallel corpus are particularly impressive, outperforming the mono-lingual baseline by 20% at 5 retrieved documents. Models estimated from the bilingual dictionary perform somewhat worse, though still outperforming mono-lingual performance at 5 documents. Both estimation methods outperform the Probabilistic Translation Model. We consider these results to be extremely encouraging, since they suggest that Cross-lingual Relevance Models perform very well in the important high-precision area.

6. CONCLUSION

We proposed a formal probabilistic model of Cross-Language Information Retrieval. The model is significantly different from other recently proposed models in that it does not attempt to translate either the query or the documents. The model starts with a query in the source language and directly estimates the model of relevant documents in the target language. Massive query expansion is an integral part of the model, rather than a heuristic addition. Our experiments demonstrate that performance of the model is as good as or better than that of previously reported models. The model performs around 90% - 95% of the strong mono-lingual baseline in terms of average precision. In terms of initial precision, the model outperforms the mono-lingual baseline by 20%. We discussed how Cross-Lingual Relevance Models can be estimated using either a parallel/comparable corpus, or a bilingual lexicon. Our experiments show that coverage is an extremely important aspect of the resources. We have been able to use a commercial MT system to increase the coverage of our corpus.

A number of questions remain open and need to be addressed in the future work. We would like to explore the use of unlabeled data to improve the coverage of the parallel corpus without performing expensive machine translation. We also plan to carry out additional experiments with sentence-aligned parallel corpora. Finally, we would like to explore applications of our model in other tasks, such as Topic Detection and Tracking, and in other languages.

7. ACKNOWLEDGMENTS

We would like to thank Jinxi Xu for providing us with the bilingual dictionary and Fang-Fang Feng for providing the Chinese segmenter and other resources. Many thanks to the reviewers, whose feedback proved to be very valuable. This work was supported in part by the Center for Intelligent Information Retrieval, and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] J. Allan, J. Callan, W. B. Croft, L. A. Ballesteros, D. Byrd, R. Swan, and J. Xu. INQUERY does battle with TREC-6. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 169–206, Gaithersburg, MD, November 1997. National Institute of Standards and Technology (NIST) and Defense Advanced Research Projects Agency (DARPA), Department of Commerce, National Institute of Standards and Technology.

[2] L. A. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In N. J. Belkin, A. D. Narasimhalu, and P. Willett, editors, *Proceedings of the Twentieth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 84–91, Philadelphia, PA, July 1997. ACM Press.

[3] L. A. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the Twenty-First Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71, Melbourne, Australia, August 1998. ACM Press.

[4] A. Berger and J. Lafferty. Information retrieval as statistical translation. In M. Hearst, F. Gey, and R. Tong, editors, *Proceedings of the Twenty-Second Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, Berkeley, CA, August 1999. ACM Press.

[5] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[6] W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors. *Proceedings of the Twenty-Fourth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, September 2001. ACM Press.

[7] J. Gao, J.-Y. Nie, J. Zhang, E. Xun, Y. Su, M. Zhou, and C. Huang. TREC-9 CLIR experiments at

	Unigrams		Bigrams		Words		Combination	
Short Queries								
Monolingual baseline	0.2447		0.2375		0.2604		0.2874	
Monolingual Relevance Model	0.2404	-2%	0.2524	+6%	0.3072	+18%	0.3100	+8%
Probabilistic Translation Model	—		—		—		0.2549	-11%
Cross-lingual Relevance Model (dictionary)	0.2776	+13%	0.2684	+13%	0.2779	+7%	0.2807	-2%
Cross-lingual Relevance Model (parallel)	0.2258	-8%	0.2519	+6%	0.2493	-4%	0.2670	-7%
Long Queries								
Monolingual baseline	0.2750		0.3090		0.2837		0.3302	
Monolingual Relevance Model	0.2835	+3%	0.3414	+10%	0.3240	+14%	0.3767	+14%
Probabilistic Translation Model	—		—		—		0.2768	-16%
Cross-lingual Relevance Model (dictionary)	0.3002	+9%	0.3074	-1%	0.3178	+12%	0.3182	-4%

Table 3: Average Precision on the TREC9 cross-language retrieval task. Cross-lingual Relevance Models perform around 95% of the strong mono-lingual baseline

Precision	Monolingual	Mono R.M.		Translation Model		R.M. (dictionary)		R.M. (parallel)	
at 5 docs	0.2880	0.3360	+17%	0.2560	-11%	0.3200	+11%	0.3520	+22%
at 10 docs	0.2600	0.2880	+11%	0.2120	-19%	0.2320	-11%	0.2880	+11%
at 15 docs	0.2240	0.2427	+8%	0.1867	-17%	0.2080	-7%	0.2453	+10%
at 20 docs	0.2120	0.2220	+5%	0.1700	-20%	0.1960	-8%	0.2080	-2%
at 30 docs	0.1867	0.1907	+2%	0.1440	-23%	0.1693	-9%	0.1827	-2%

Table 4: Initial precision on the TREC9 CLIR task. Cross-lingual Relevance Models noticeably outperform the mono-lingual baselines.

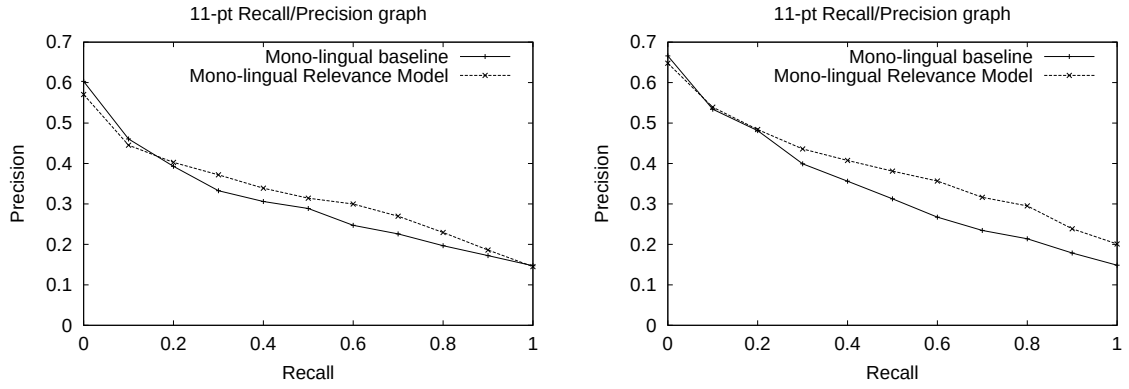

Figure 3: Mono-lingual results. Relevance Models provide a higher baseline for both short queries (left) and long queries (right).

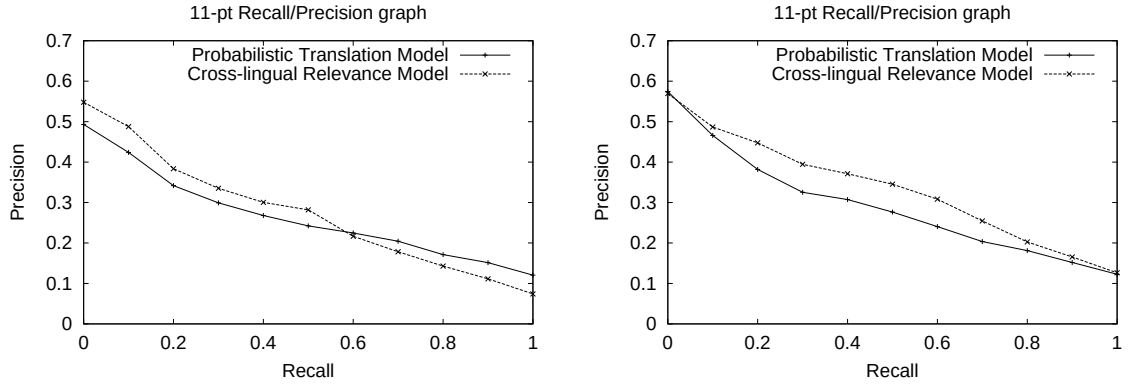

Figure 4: Cross-lingual Relevance Models outperform the Probabilistic Translation Model on both the short (left) and long (right) queries.

MSRCN. In Voorhees and Harman [17], pages 343–354.

[8] T. Gollins and M. Sanderson. Improving cross language information retrieval with triangulated translation. In Croft et al. [6], pages 90–95.

[9] D. Hiemstra and F. de Jong. Disambiguation strategies for cross-language information retrieval. In S. Abiteboul and A.-M. Vercoustre, editors, *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libaries, ECDL'99*, volume 1696 of *Lecture Notes in Computer Science*, pages 274–293. Springer-Verlag, Paris, September 1999.

[10] D. Hiemstra, F. de Jong, and W. Kraaij. A domain specific lexicon acquisition tool for cross-language information retrieval. In L. Devroye and C. Chrisment, editors, *Proceedings of the Fifth RIAO International Conference*, pages 255–270, Montréal, Canada, 1997. Centre de Hautes Études Internationales d'Informatique Documentaire (C.I.D).

[11] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In Croft et al. [6], pages 111–119.

[12] V. Lavrenko and W. B. Croft. Relevance-based language models. In Croft et al. [6], pages 120–127.

[13] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[14] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304, 1977. Reprinted in [16].

[15] P. Sheridan and J. P. Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of the Nineteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 58–64, Zurich, Switzerland, August 1996. ACM Press.

[16] K. Sparck Jones and P. Willett, editors. *Readings in information retrieval*. Multimedia Information and Systems. Morgan Kaufmann, San Francisco, CA, 1997.

[17] E. M. Voorhees and D. K. Harman, editors. *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, Gaithersburg, MD, November 2000. Department of Commerce, National Institute of Standards and Technology.

[18] J. Xu and R. Weischedel. TREC-9 cross-lingual retrieval at BBN. In Voorhees and Harman [17], pages 106–116.

[19] J. Xu, R. Weischedel, and C. Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In Croft et al. [6], pages 105–110.