A Maximum Coherence Model for Dictionary-based Cross-language Information Retrieval

Yi Liu
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
liuyi3@cse.msu.edu

Rong Jin
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
rongjin@cse.msu.edu

Joyce Y. Chai
Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
jchai@cse.msu.edu

ABSTRACT

One key to cross-language information retrieval is how to efficiently resolve the translation ambiguity of queries given their short length. This problem is even more challenging when only bilingual dictionaries are available, which is the focus of this paper. In the previous research of cross-language information retrieval using bilingual dictionaries, the word co-occurrence statistics is used to determine the most likely translations of queries. In this paper, we propose a novel statistical model, named "maximum coherence model", which estimates the translation probabilities of query words that are consistent with the word co-occurrence statistics. Unlike the previous work, where a binary decision is made for the selection of translations, the new model maintains the uncertainty in translating query words when their sense ambiguity is difficult to resolve. Furthermore, this new model is able to estimate translations of multiple query words simultaneously. This is in contrast to many previous approaches where translations of individual query words are determined independently. Empirical studies with TREC datasets have shown that the maximum coherence model achieves a relative 10% - 40% improvement in cross-language information retrieval, comparing to other approaches that also use word co-occurrence statistics for sense disambiguation.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Experimentation

Keywords

maximum coherence model, co-occurrence statistics, cross-language information retrieval

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

1. INTRODUCTION

To overcome the language barrier in cross-language information retrieval (CLIR), either queries or documents are translated into the language of their counterparts. Usually it is simpler and more efficient to translate queries than to translate documents because queries are generally much shorter than documents. Given its short length, how to disambiguate translations of a query has become a challenging problem in cross-language information retrieval. Many approaches, such as statistical translation models [7, 11, 22] and relevance language models [12, 13, 14, 17], rely on parallel bilingual corpora for query translation disambiguation. Often, they learn an association between the words in the language of queries and the language of documents from a bilingual corpus, and apply the association to disambiguate translations of queries. However, it is usually not only time consuming but also expensive to acquire large parallel bilingual corpora, particularly for minor languages. Due to the increasing availability of machine readable dictionaries, much of the research effort in CLIR has been put into the dictionary-based approaches.

The simplest approach toward dictionary-based CLIR is to use all the translations of query words provided by the dictionary equally [5, 6]. This amounts to no sense disambiguation for query words. Other approaches try to resolve the translation ambiguity by measuring the *coherence* of a translation word to the entire query. Typically, the coherence score of a translation word is computed using word co-occurrence statistics. Given a query, a translation of a query word is assigned with a high coherence score when it co-occurs frequently with the translations of other query words. The selection of translation words are then determined by their coherence scores: in approaches [2, 5, 6, 8, 10, 12], for each query word, only the translation with the highest coherence score is selected; in approaches [15, 16], a translation word is selected when its coherence score exceeds a certain threshold. In both approaches, a *selection* strategy is used, namely for each query word, a *binary* decision has to be made as to which translation(s) of the word should be used for the translation of the query. Given the short length of queries and the large variance existing in mapping information across different languages, such binary decisions are usually difficult, if not impossible, to make. We call this problem "*translation uncertainty problem*". Another problem with the selection-based approaches is that the transla-

tion of one query word is usually determined independently from the translations of others, which we call "*translation independence assumption*". This assumption is reflected in the calculation of coherence scores. Usually, the coherence score of a translation word is computed as the sum of similarities to all the translations of query words provided by the dictionary. As a result, coherence scores are estimated independently from the choice of translations for query words, which leads the selection of translations for different query words to be independent.

In this paper, we propose a novel statistical model, named "**maximum coherence model**". It estimates the translation probabilities of query words by maximizing the overall coherence of the corresponding query, which we call "*maximum coherence principle*". In particular, the proposed model explicitly addresses the two problems mentioned above: to resolve the translation uncertainty problem, the maximum coherence model maintains the uncertainty in translating queries through the estimation of translation probabilities for query words; to drop the translation independence assumption, the new model estimates the translation probabilities for all query words simultaneously. To speed up the computation, a quadratic programming technique is employed to efficiently solve the related optimization problem. Our empirical studies with TREC datasets have shown that the maximum coherence model outperforms the selection-based approaches with relative improvements ranging from 10% to 40%.

The rest of the paper is structured as follows: Section 2 briefly reviews the related work in selection-based approaches for query translation disambiguation. Section 3 describes our maximum coherence model, and the procedure for solving the related optimization problem. Section 4 presents the experimental results. Section 5 concludes this work.

2. RELATED WORK

One of the major factors that can potentially degrade the effectiveness of dictionary-based cross-language information retrieval is the ambiguity in translating query words [3, 8]. In the efforts to resolve this translation ambiguity, several recent studies [2, 5, 6, 8, 10, 12, 15, 16] have suggested the strategy of translation selection by exploiting word co-occurrence patterns. Usually a similarity measurement between two translation words is defined in the form of word co-occurrence statistics. With the word similarities, we can then measure the coherence of a translation word with regard to a query. Only translation words with high coherence scores will be selected for the translation of the query.

Ideally, for each query word, we should select the translation(s) that is consistent with the selected translations for other query words. Apparently, this becomes a "chicken-egg" problem since the selection of translations for one word is determined by the translations selected for other words. Thus, due to the computational concern, most selection-based approaches [1, 8, 9] adopted an approximate solution. For each query word, it selects the translation that is most consistent with *all* the translations provided by the dictionary for all query words, including both the selected and the unselected translations. Typically, a translation selection strategy can be formulated into the following algorithm:

Approximate Translation Selection Algorithm

1. Given a query $\mathbf{q}^s = \{q_1^s, q_2^s, \cdots, q_{m^s}^s\}$ in the source language , for each query word q_i^s, look up the dictionary for the set of all translation $S_i = \{w_{i,j}^t\}$
2. For each set S_i
   (a) For each translation $w_{i,j}^t$ in S_i, define the similarity measurement between the word $w_{i,j}^t$ and the set $S_{i'}(i' \neq i)$ as the sum of the similarities between $w_{i,j}^t$ and each word in the set $S_{i'}$, i.e.,
   $$sim(w_{i,j}^t, S_{i'}) = \sum_{\forall w_{i',l}^t \in S_{i'}} sim(w_{i,j}^t, w_{i',l}^t) \quad (1)$$
   (b) Compute the coherence score for $w_{i,j}^t$ as
   $$f(w_{i,j}^t) = \sum_{\forall i' \neq i} sim(w_{i,j}^t, S_{i'}) \quad (2)$$
   (c) Select the word q_i^t in S_i with the highest coherence score
   $$q_i^t = \arg\max_{w_{i,j}^t} f(w_{i,j}^t) \quad (3)$$

The definition of similarity between two words in the above algorithm can take various forms of co-occurrence statistics, such as Dice similarity (as in [1]), mutual information (as in [15, 16]) or its variants (as in [8, 9]). In addition to selecting the most likely translation for each query word, other selection-based approaches have been tried, such as selecting the best N translations [6] and selecting translations by a predefined threshold [15, 16].

Apparently the above approximate algorithm is not ideal. In particular, the coherence score for a translation is computed with regard to both selected and unselected translations. As a result of such an approximation, translation of different query words are determined independently, which leads to the translation independence problem as discussed in the introduction section. In the proposed model, by formulating the problem of translation selection into a quadratic programming problem, we are able to efficiently estimate the translations for *all* query words *simultaneously*. Furthermore, in contrast to the selection-based approaches that make binary decision for each translation, the new model employs soft probabilities for representing both selected and unselected translations. This is particularly useful when binary decisions are hard to make, for instance, all the translations of a query word have very similar coherence scores.

3. MAXIMUM COHERENCE MODEL

The essential idea of the maximum coherence model is to learn a set of translation probabilities for query words from word co-occurrence statistics that maximizes the overall coherence of the corresponding query. This is referred as "maximum coherence principle". In the following subsections, we will first describe the proposed statistical model and the definition of the overall coherence for a query, followed by a description of the procedure that solves the related optimization problem efficiently.

Before starting the discussion of the proposed model, we would like to introduce the notations that is used throughout this paper. Similar to other CLIR papers, "source language"

refers to the language of queries, and "target language" refers to the language of documents. In order to differentiate the source language from the target language, a superscript s is used for any variable related to the source language and a superscript t is used for any variable related to the target language. Let a query of the source language be denoted by $\mathbf{q}^s = \{q_1^s, q_2^s, \cdots, q_{m^s}^s\}$, where m^s is the number of distinct words in $\mathbf{q}^s$. Let m^t be the total number of distinct translations provided by the dictionary for all the words in query $\mathbf{q}^s$. Let matrix $\mathbf{T}$ represent the part of the bilingual dictionary related to query $\mathbf{q}^s$, i.e., $\mathbf{T} = [t_{k,j}]_{m^s \times m^t}$. An element $t_{k,j}$ in $\mathbf{T}$ is 1 if the j-th word of the target language appears as a translation in the dictionary for the k-th word in the source language and 0 otherwise. Also we use $\mathbf{r}_k$ to denote the set of translations provided by the dictionary for a word w_k^s in the query $\mathbf{q}^s$.

3.1 Modelling the Uncertainty in Query Translation

To address the problem of translation uncertainty, the new model introduces translation probabilities to capture the uncertainty in translating queries.

Let $p_{k,j}$ denote the probability of translating a word w_k^s in the source language into a word w_j^t in the target language, given the context of query $\mathbf{q}^s$. It is defined as

$$p_{k,j} = \Pr(w_j^t | w_k^s, \mathbf{q}^s) \quad (4)$$

which satisfies

$$\sum_{\forall j, w_j^s \in \mathbf{r}_k} p_{k,j} = 1$$
$$p_{k,j} \geq 0$$

By aggregating the translation probabilities for all the words in $\mathbf{q}^s$, we can define a matrix for translation probabilities:

$$\mathbf{P} = [p_{k,j}]_{m^s \times m^t}$$

With the translation probabilities $p_{k,j}$, we can now define a statistical retrieval model for CLIR [11, 13]. In particular, we estimate $\Pr(\mathbf{d}^t|\mathbf{q}^s)$, i.e., the probability for a document $\mathbf{d}^t$ in the target language to be relevant to a query $\mathbf{q}^s$ in the source language. By the Bayes' law, this probability can be approximated as

$$\Pr(\mathbf{d}^t|\mathbf{q}^s) = \frac{\Pr(\mathbf{q}^s|\mathbf{d}^t) \cdot \Pr(\mathbf{d}^t)}{\Pr(\mathbf{q}^s)} \sim \Pr(\mathbf{q}^s|\mathbf{d}^t)$$

The last step is based on the assumption that document prior $\Pr(\mathbf{d}^t)$ follows a uniform distribution.

Taking the logarithm of $\Pr(\mathbf{d}^t|\mathbf{q}^s)$, we have

$$\begin{aligned} \log \Pr(\mathbf{d}^t|\mathbf{q}^s) &\sim \log \Pr(\mathbf{q}^s|\mathbf{d}^t) \\ &\sim \sum_{w^t} \Pr(w^t|\mathbf{q}^s) \log \Pr(w^t|\mathbf{d}^t) \\ &= \sum_{w^t} \sum_{w^s} \Pr(w^t|w^s) \Pr(w^s|\mathbf{q}^s) \log \Pr(w^t|\mathbf{d}^t) \end{aligned} \quad (5)$$

Here $\Pr(w^t|\mathbf{d}^t)$ is a monolingual language model for document $\mathbf{d}^t$ in the target language; $\Pr(w^t|w^s)$ is the probability for translating query word w^s into w^t; and $\Pr(w^s|\mathbf{q}^s)$ is a monolingual language model for query $\mathbf{q}^s$ in the source language, which can also be seen as the weight assigned to the query word w^s. For the sake of simplicity, we assume equal weights for all query words in the source language .

3.2 Maximum Coherence Model

The crucial part of our model is to determine the translation probabilities for a given query. To this end, we propose the maximum coherence model that automatically learns translation probabilities for query words from word co-occurrence statistics. The key of this learning procedure is to first define the overall coherence for a query, and then efficiently identify the set of translation probabilities that maximizes the overall coherence measurement. Using the translation probabilities introduced in the previous subsection, we can now define a probabilistic measurement for the overall coherence for a query $\mathbf{q}^s$, i.e.,

$$Co(\mathbf{q}^s; \mathbf{T}) = \sum_{\substack{\forall k, w_k^s \in \mathbf{q}^s \\ \forall k', w_{k'}^s \in \mathbf{q}^s}} \sum_{\substack{\forall j, w_j^t \in \mathbf{r}_k \\ \forall j', w_{j'}^t \in \mathbf{r}_{k'}}} p_{k,j} \cdot s_{j,j'}^t \cdot p_{k',j'} \quad (6)$$

where $s_{j,j'}$ is a similarity measurement between word w_j^t and $w_{j'}^t$ that can be derived from word co-occurrence statistics. In the previous studies of selection-based approaches, several metrics have been used for similarity $s_{j,j'}$, including the mutual information [8, 16], or the Dice similarity [1, 2]. In this paper, we adopt the mutual information metric for similarity measurement, which is defined as

$$\begin{aligned} s_{j,j'}^t &= MI(w_j^t, w_{j'}^t) \\ &= \Pr(w_j^t, w_{j'}^t) \times \log \frac{\Pr(w_j^t, w_{j'}^t)}{\Pr(w_j^t) \times \Pr(w_{j'}^t)} \end{aligned} \quad (7)$$

$\Pr(w_j^t)$ is the unigram probability for word w_j^t, and $\Pr(w_j^t, w_{j'}^t)$ is the joint probabilities for word w_j^t and $w_{j'}^t$ to co-occur in same documents. Both probabilities can be acquired by simply counting the term frequency of single words and the frequency of co-occurrence between two words.

Using the matrix notation, the expression for the overall coherence can be simplified as

$$Co(\mathbf{q}^s; \mathbf{T}) = \mathbf{e}^T \mathbf{P} \mathbf{S} \mathbf{P}^T \mathbf{e} \quad (8)$$

where $\mathbf{e} = [1, 1, \ldots, 1]^T$ and $\mathbf{S} = [s_{j,j'}^t]_{m^t \times m^t}$. Based on the principle of maximum coherence, the optimal set of translation probabilities is acquired by maximizing the overall coherence $Co(\mathbf{q}^s; \mathbf{T})$, i.e.,

$$\max_{\mathbf{P}} \mathbf{e}^T \mathbf{P} \mathbf{S} \mathbf{P}^T \mathbf{e} \quad (9)$$
$$s.t. \sum_{\forall j, w_j^t \in \mathbf{r}_k} p_{k,j} = 1 \quad \text{for all } k$$
$$p_{k,j} \geq 0 \quad \text{for all } k$$

To avoid unstable results, similar to logistic regression [18] and support vector machine [4], a regularizer is introduced into the above objective function, which is expressed as

$$Trace(\mathbf{P}\mathbf{P}^T\mathbf{1}) \quad (10)$$

where $\mathbf{1}$ is a matrix whose elements are all 1. $Trace(\mathbf{P}\mathbf{P}^T\mathbf{1}) = \sum_{k,k'} \sum_j p_{k,j} p_{k',j}$, i.e., the sum of all elements in $\mathbf{P}\mathbf{P}^T$. Similar to the uninformative priors used in the Bayesian learning, the goal of this regularizer is to reflect our prior knowledge of translation probabilities — without context, we assume that all translations provided by a bilingual dictionary are equally likely to be selected. By combining the regularizer with the coherence measurement, we now have a

regularized optimization problem, i.e.,

$$\max_{\mathbf{P}} \mathbf{e}^T\mathbf{P}\mathbf{S}\mathbf{P}^T\mathbf{e} - C_p \cdot Trace(\mathbf{P}\mathbf{P}^T\mathbf{1}) \quad (11)$$

$$s.t. \sum_{\forall j, w_j^t \in \mathbf{r}_k} p_{k,j} = 1 \quad \text{for all } k$$

$$p_{k,j} \geq 0 \quad \text{for all } k$$

where C_p is a constant that balances the contribution between the coherence measurement and the regularizer. It is determined empirically in our experiments. Note that, in the above formalization, the translation probabilities for all query words are estimated simultaneously through the computation of $\mathbf{P}$. This is in contrast to the selection-based approaches, in which the selection of translations for individual query words are determined independently.

3.3 Solving the Optimization Problem

The optimization problem in (11) is in fact a standard quadratic programming (QP) problem [19]. To write it in an explicit QP form, we define

$$\mathbf{P}_{m^s \times m^t} = \begin{pmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \\ \vdots \\ \mathbf{p}_{m^s}^T \end{pmatrix} \qquad \mathbf{T}_{m^s \times m^t} = \begin{pmatrix} \mathbf{t}_1^T \\ \mathbf{t}_2^T \\ \vdots \\ \mathbf{t}_{m^s}^T \end{pmatrix} \quad (12)$$

$$\mathbf{q}_{m^s m^t \times 1} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_{m^s} \end{pmatrix} \qquad \overline{\mathbf{q}}_{m^s m^t \times 1} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \vdots \\ \mathbf{t}_{m^s} \end{pmatrix} \quad (13)$$

$$\mathbf{A}_{m^s m^t \times m^s m^t} = \mathbf{1} \otimes \mathbf{S} \quad (14)$$

$$\mathbf{E}_{m^s \times m^s m^t} = diag(\mathbf{t}_1^T, \mathbf{t}_2^T, \ldots, \mathbf{t}_{m^s}^T) \quad (15)$$

$$\mathbf{H} = \mathbf{A} - C_p \mathbf{1}_{m^s \times m^s} \otimes \mathbf{I}_{m^t \times m^t} \quad (16)$$

where $\otimes$ represents *kronecker product*. It is easy to derive that

$$\mathbf{e}^T\mathbf{P}\mathbf{S}\mathbf{P}^T\mathbf{e} = \mathbf{q}^T\mathbf{A}\mathbf{q} \quad (17)$$

$$Trace(\mathbf{P}\mathbf{P}^T\mathbf{1}) = \mathbf{q}^T(\mathbf{1}_{m^s \times m^s} \otimes \mathbf{I}_{m^t \times m^t})\mathbf{q} \quad (18)$$

Using Equation (12) - (18), the optimization problem in (11) can be rewritten in a standard form of the QP problem

$$\max_{\mathbf{q}} \mathbf{q}^T\mathbf{H}\mathbf{q} \quad (19)$$

$$s.t. \quad \mathbf{E}\mathbf{q} = \mathbf{1}_{m^s \times 1} \quad (20)$$

$$0 \leq \mathbf{q} \leq \overline{\mathbf{q}} \quad (21)$$

The quadratic programming is a well studied optimization problem and can be solved efficiently. In our experiment, we use the QP package in MATLAB [21].

3.4 Computational Concerns

For the QP problem formulated in (19), the problem size appears to be large because the number of variables in vector $\mathbf{q}$ is $m^s m^t$, i.e., the product between the number of unique query words and the number of distinct translation words provided by the dictionary. But, notice that in the constraint (21), $\overline{\mathbf{q}}$, i.e., the upper bound of translation probabilities, is a concatenation of translation vectors $\mathbf{t}_i$ obtained from dictionary $\mathbf{T}$. Given that most query words only have a few translations, most of the elements in the bilingual dictionary $\mathbf{T}$ will be zeros. As a result, most elements in the upper bound vector $\overline{\mathbf{q}}$ are zeros, which leads to the zero values for the corresponding translation probabilities in $\mathbf{q}$. Hence, the number of non-zero translation probabilities in $\mathbf{q}$ is no more than the total number of translations provided by the bilingual dictionary for the query words, which is usually much smaller than the product $m^s m^t$. Thus, the computation cost of the maximum coherence model is modest for real CLIR practice, if not overestimated.

4. EXPERIMENTS AND DISCUSSIONS

The goal of this experiment is to examine the effectiveness of the proposed model for cross-language information retrieval. In particular, three research questions will be addressed in this empirical study:

1. *Is the proposed maximum coherence model effective for cross-language information retrieval*? To obtain a comprehensive view, we compare the maximum coherence model to existing selection-based approaches using different types of queries and documents.

2. *How important is it for a query disambiguation algorithm to include translation uncertainty in its analysis*? To address this question, we will examine the importance of including translation uncertainty in cross-language information retrieval through case studies.

3. *How important is it to remove the translation independence assumption for cross-language information retrieval* ? To address this question, we will examine the impact of the translation independence assumption on cross-language information retrieval through case studies.

4.1 Experiment Setup

All our experiments are retrieval of English documents using Chinese queries. The document collections used in this experiment are from TREC ad hoc test collections, including

AP88-89 164,835 documents from Associated Press(1988, 1989)

WSJ87-88 83,480 documents from Wall Street Journal (1987, 1988)

DOE1-2 226,087 documents from Department of Energy abstracts [1]

In addition to the homogeneous collections listed above, we also tested the proposed model against heterogeneous collections that are formed by combining multiple homogeneous collections. In particular, two heterogeneous collections are created: collection AP88-89 + WSJ87-88, and collection AP89 + WSJ87-88 + DOE1-2. In a heterogeneous collection, words are more likely to carry multiple senses than words from a homogeneous collection, which will increase the difficulty for an automatic algorithm to disambiguate the senses of query words using the pairwise word similarities. The SMART system [20] is used to process document collections. Each document is first parsed into tokens

[1]DOE1-2 collection is not used as one of the homogeneous datasets in our experiments because DOE1-2 collection provides no relevant documents for a majority of the queries used in this experiment. It is only used to create heterogeneous collections by combining with the other two homogeneous collections.

with stop words removed, and then tokens are stemmed using the Porter algorithm. Finally, each document is represented as a bag of stemmed words. Since our goal is to illustrate the advantage of the proposed statistical model, we did not apply more sophisticated procedures for text analysis in our experiment, such as phrase identification.

Our queries come from a manual Chinese translation of TREC-3 ad hoc topics (topic 151-200). To fully examine the effectiveness of the proposed model, we test it against both the long Chinese queries and the short Chinese queries. A short Chinese query is created by translating the "title" field of an English query into Chinese; a long Chinese query is formed by combining the Chinese translations of both the "title" field and the "description" field in an English query. The average length of short Chinese queries is 9.64 Chinese characters, and 30.72 Chinese characters for long queries. Since most of its words in a short query are highly relevant to the topic of the query, we would expect that query disambiguation approaches based on word similarities will work well. In contrast, a long query usually include words either irrelevant or only slightly relevant to its topic. As a result, even a translation word that is coherent with the translations of many query words may not necessarily be a good candidate for selection. Hence, a long query usually poses a more challenging problem than a short query for a translation disambiguation algorithm based on word similarity information. Finally, the relevance judgments for the original English queries are used as the relevance judgments for their Chinese translations.

The Chinese-English dictionary used in our experiments comes from Linguistic Data Consortium (LDC, http://www.-ldc.upenn.edu), which consists of translations for 53061 Chinese words. Since our experiments do not involve the processing of English phrases, for any English phrase that is the translation of a Chinese word, we simply treat it as a bag of words.

To evaluate the effectiveness of the proposed method, we implement two baseline models that use translation selection methods. The first baseline model selects the most likely translation for each query word, which we call "BESTONE". The details of this model has been described in section of related work. The second model, which we call "ALLTRANS", makes no efforts for translation disambiguation by simply including all the translations provided by the dictionary for query words into the final query translation. Finally, for easy reference, we use the abbreviation "MAXCO" for our maximum coherence model. The constant C_p for the regularizer is set to be $\frac{4}{(m^t)^2}\sum_{j,j'} s^t_{j,j'}$ based on our empirical experience.

4.2 Comparison to Selection-based Approaches

Table 1 lists the average precision across 11 recall points for both the homogeneous collections and the heterogeneous collections. As indicated in Table 1 the proposed model (i.e., "MAXCO") is able to outperform the two baseline models for both short queries and long queries across all four different collections. Furthermore, we plot the precision-recall curves for both the short queries and the long queries in Figure 1 and Figure 2, respectively. As illustrated in Figure 1 and 2, for all four collections, the precision-recall curves of the maximum coherence model always stay above the curves of the other two models. Based on these results, we conclude that the maximum coherence model performs substantially better than the other two selection-based approaches for cross-language information retrieval.

A further examination of results in Table 1 gives rise to the following observations:

1. In general, the retrieval accuracy for heterogeneous collections appears to be worse than that for homogeneous collections. In particular, a substantial decrease in the average precision is observed for all three methods when the collection of DOE1-2 is included in the heterogeneous collection. This result is in accordance with our previous analysis, i.e., words from heterogeneous collections are more likely to have multiple senses, thus resulting in higher translation ambiguity.

2. A better retrieval is achieved for short queries than for long queries. The degradation in performance between long queries and short queries is more significant for heterogeneous collections than for homogeneous collections. Again, this is consistent with our previous analysis: long queries are usually more difficult to disambiguate for algorithms based on word similarities. Despite of the general belief in monolingual IR that long queries are less ambiguous than short ones, long queries are generally more challenging for translation disambiguation. This is because long queries tend to include more words that are either irrelevant or only slightly relevant to their topics, which makes the estimation of coherence scores for translation candidates unreliable. In fact, among the three methods, the simple method "ALLTRANS" appears to be relatively more robust than the other two. This is because the "ALLTRANS" method does not employ any strategy for query translation disambiguation.

3. The "BESTONE" method does not consistently outperform the "ALLTRANS" method. In fact, for the long queries, the "ALLTRANS" method performs better than the "BESTONE" method across full different collections. Similar to the previous analysis, this phenomenon can be attributed to the fact that long queries are rather noisy and likely to include irrelevant words. This result indicates that the the "BESTONE" method can be sensitive to the noises present in queries. Given that a significant amount of noise can be present in queries, it is important to maintain the uncertainty of translation in the retrieval process. Note that our results appear to be inconsistent with the finding in [8]. However, the setup of our experiments is rather different from theirs. For example, we did not identify English phrases in our text processing, which have shown to be an important factor in CLIR [3, 8]. Although phrase analysis is important to CLIR, a generic probabilistic model is beneficial to CLIR of any languages, particularly when linguistic resources are scarce.

4.3 The Necessity of Including Translation Uncertainty

To demonstrate the uncertainty in query translation, in Figure 3, we list the translation probabilities for three Chinese words that are estimated by the maximum coherence model. As we can see, a significant variance exists in the distribution of translation probabilities across different Chinese words. The first example in the figure shows an almost

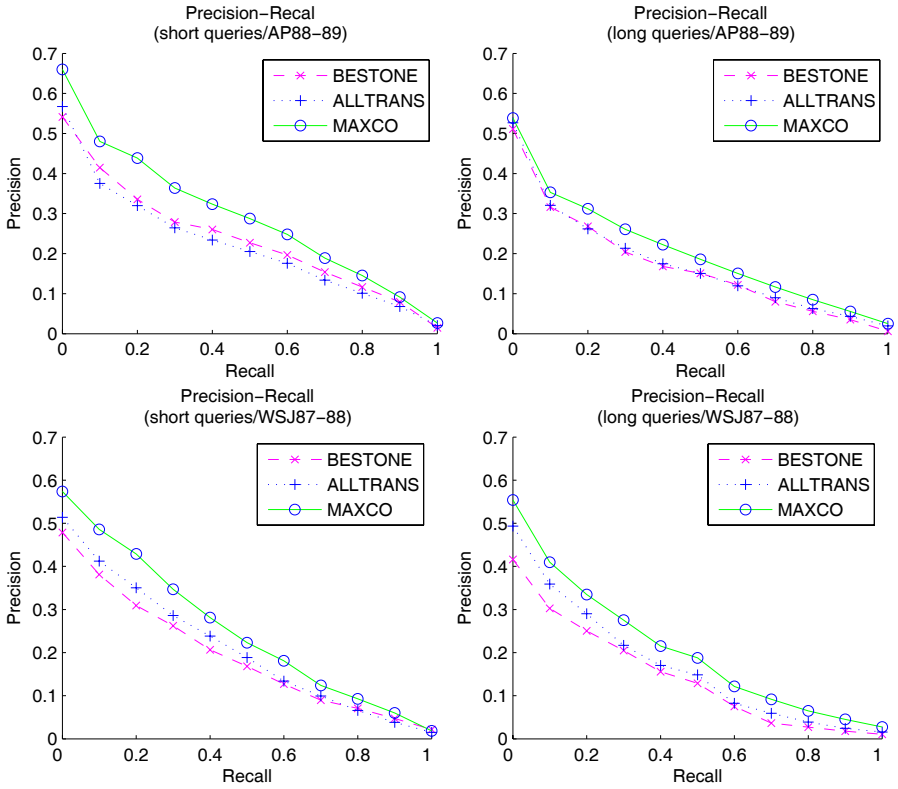

Figure 1: Comparison of CLIR performance on homogeneous datasets using both short and long queries. The upper two figures are for AP88-89 dataset, and the lower two are for WSJ87-88 dataset. The left two figures are for short queries, and the right two are for long queries.

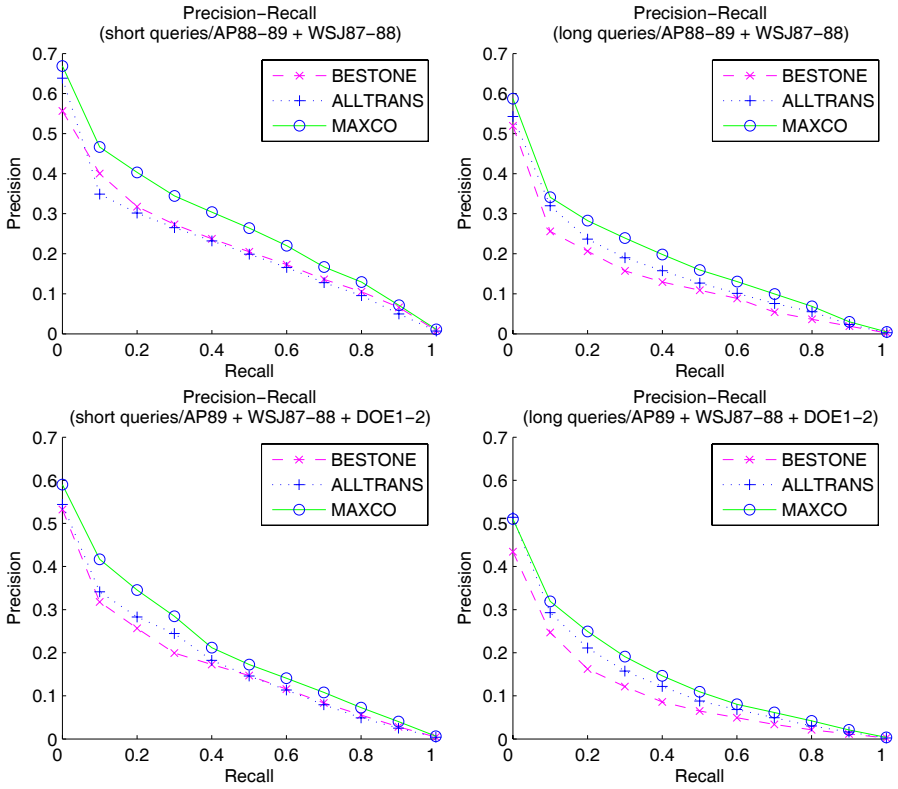

Figure 2: Comparison of CLIR performance on heterogeneous datasets using both short and long queries. The upper two figures are for AP88-89 + WSJ87-88, and the lower two are for AP89 + WSJ87-88 + DOE1-2 dataset. The left two figures are for short queries, and the right two are for long queries.

Table 1: 11-point average precision for both short and long queries on TREC datasets
(The last two columns list the relative improvements of our maximum coherence model over the other two methods)

	BESTONE	ALLTRANS	MAXCO	(M-B)/B	(M-A)/A
Short Queries					
AP88-89	0.2381	0.2241	0.2959	+24.28%	+32.04%
WSJ87-88	0.1966	0.2129	0.2560	+30.21%	+20.24%
AP88-89 + WSJ87-88	0.2253	0.2209	0.2772	+23.04%	+25.49%
AP89 + WSJ87-88 + DOE1-2	0.1739	0.1829	0.2172	+24.90%	+18.75%
Long Queries					
AP88-89	0.1749	0.1803	0.2096	+19.84%	+16.25%
WSJ87-88	0.1478	0.1727	0.2116	+43.17%	+22.52%
AP88-89 + WSJ87-88	0.1433	0.1665	0.1947	+35.87%	+16.94%
AP89 + WSJ87-88 + DOE1-2	0.1122	0.1411	0.1576	+40.46%	+11.69%

	upheaval	commotion	turbulence	unrest	turmoil
动荡	0.19931	0.19723	0.19882	0.20209	0.20255

	intent	spring	motive	inducement	incentive
动机	0.22821	0.19645	0.30384	0.13196	0.13954

	acid	sour	sore	ache
酸	0.79070	0.06422	0.07552	0.06956

Figure 3: Examples of translation probabilities estimated by the maximum coherence model.

CH Term	EN Translation	Selection (BESTONE)	Trans. Prob. (MAXCO)
独立	independent		0.36208
	on one's *own*		0.24481
	stand *alone*	x	0.39311
出版	press	x	0.33789
	publish		0.19973
	put *out*		0.33311
	print		0.12927
消失	disappear	x	0.34941
	demise		0.32035
	fade		0.33024
征兆	symptom		0.16698
	omen		0.16155
	sign	x	0.35016
	portent		0.16097
	premonition		0.16034

Original TREC Topic in English (topic 187 'title' field):
Signs of the Demise of Independent Publishing

Figure 4: An example of query translation, using the "BESTONE" method and the maximum coherence model. (English words in italicized font are removed as stop words.)

uniform distribution over all translations, while the third one illustrates a very skewed distribution. Meanwhile, the second example provides a distribution that is neither uniform nor totally skewed. These three examples illustrate the "*translation uncertainty problem*", which we have addressed in previous sections. Furthermore, the diversity in the distribution of translation probabilities makes it difficult for a selection-based approach to perform well over all different cases. For example, the "BESTONE" method is able to work well for the third example but will fail in the first one. On the other hand, the "ALLTRANS" method would be perfect for the first example but not for the third one. Base on the above analysis, we conclude that it is important to capture the translation uncertainty and the diversity of translation uncertainty in a probabilistic model.

4.4 The Impact of Translation Independence Assumption on Query Disambiguation

To illustrate the impact of the translation independence assumption on query translation disambiguation, consider the example in Figure 4. This query consists of four Chinese words, and the English translations for each Chinese are provided by the dictionary are listed in the second column. For the purpose of illustration, the original English query is also included at the bottom of the figure. The English translations selected by the "BESTONE" method are listed in the third column, marked by small crosses. The translation probabilities from Chinese words to their English translations estimated by the maximum coherence model are listed in the last column.

Comparing to the original English query, we see that the "BESTONE" method makes incorrect translation selection for both the first and the second Chinese words. For the first one, the correct English translation should be "independent", instead of "stand (alone)" [2]. The better translation for the second Chinese word should be "publish" instead of "press". One reason for such mistakes is that in the "BESTONE" method, the coherence score of a translation is computed based on all the English translations provided by the dictionary for the Chinese words in the query. Thus, the coherence score of one translation word is completely independent from the selection of other translations. Since both "stand" and "press" are common in English, their overall coherence scores turn out to be larger than the coherence scores of other words, which lead them to be selected by the "BESTONE" method. In contrast, in the maximum coherence model, the estimation of translation probabilities for one word is dependent on the estimation of translation probabilities for other words. As a result, the maximum coherence model is able to adjust the mistakes by assigning significant amounts of probability mass to the correct translations. For example, for the first Chinese word, the maximum coherence model is able to assign a probability to the correct English translation "independent" comparable to the probability assigned to the translation "stand (alone)".

[2] "alone" is removed as a stop word and does not count in the translation. It is listed only for clarity purpose.

5. CONCLUSIONS

In this paper, we propose a novel statistical model for cross-language information retrieval, named "maximum coherence model". It utilizes word co-occurrence statistics for estimating translation probabilities that are effective for query disambiguation. Compared to the selection-based approaches, the merits of the maximum coherence model are twofold: 1) It preserves the translation uncertainty through the estimation of translation probabilities; 2) It estimates the translations for all query words simultaneously. Empirical results under various scenarios have shown that the proposed model is able to perform substantially better than the existing selection-based approaches.

In the future, we plan to improve the robustness of the maximum coherence model with regard to the query noises, which has led to significant degradation in the retrieval accuracy in our experiments.

6. REFERENCES

[1] M. Adriani. Dictionary-based clir for the clef multilingual track. In *Working Notes of the Workshop in Cross-Language Evaluation Forum (CLEF)*, Lisbon, September 2000.

[2] M. Adriani. Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Inf. Retr.*, 2(1):71–82, 2000.

[3] L. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 84–91. ACM Press, 1997.

[4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel, editors. *Different Approaches to Cross Language Information Retrieval*, number 37 in Language and Computers: Studies in Practical Linguistics, Amsterdam, 2001. Rodopi.

[6] M. W. Davis. New experiments in cross-language text retrieval at NMSU's computing research lab. In D. K. Harman, editor, *The Fifth Text REtrieval Conference (TREC-5)*. NIST, 1996.

[7] M. Federico and N. Bertoldi. Statistical cross-language information retrieval using N-best query translations. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 167–174. ACM Press, 2002.

[8] J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. Improving query translation for cross-language information retrieval using statistical models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–104. ACM Press, 2001.

[9] J. Gao, M. Zhou, J.-Y. Nie, H. He, and W. Chen. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 183–190. ACM Press, 2002.

[10] D. A. Hull and G. Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57. ACM Press, 1996.

[11] W. Kraaij, J.-Y. Nie, and M. Simard. Embedding web-based statistical translation models in cross-language information retrieval. *Comput. Linguist.*, 29(3):381–419, 2003.

[12] W. Kraaij, R. Pohlmann, and D. Hiemstra. Twenty-one at TREC-8: using language technology for information retrieval. In E. M. Voorhees and D. K. Harman, editors, *The Eigth Text REtrieval Conference (TREC-8)*, volume 8, pages 285–300. National Institute of Standards and Technology, NIST, 2000. NIST Special Publication 500-246.

[13] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 175–182. ACM Press, 2002.

[14] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM Press, 2001.

[15] A. Maeda, F. Sadat, M. Yoshikawa, and S. Uemura. Query term disambiguation for web cross-language information retrieval using a search engine. In *IRAL '00: Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 25–32. ACM Press, 2000.

[16] S. H. M. Myung-Gil Jang and S. Y. Park. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the association for computational linguistics*, 1999.

[17] J.-Y. Nie and M. Simard. Using statistical translation models for bilingual ir. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 137–150. Springer-Verlag, 2002.

[18] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[19] W. M. P.E. Gill and M. Wright. *Practical Optimization*. Academic Press, Inc., San Diego, USA, 1981.

[20] G. Salton, editor. *The SMART retrieval system*. Prentice-Hall, 1971.

[21] The Mathworks. http://www.mathworks.com/.

[22] J. Xu and R. Weischedel. TREC-9 cross-lingual retrieval at BBN. In *Proceedings of the TREC-9 Conference*, 2001.