

Query Translation Disambiguation As Graph Partitioning

Yi Liu and Rong Jin

Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, U.S.A.
{liuyi3, rongjin}@cse.msu.edu

Abstract

Resolving ambiguity in the process of query translation is crucial to cross-language information retrieval when only a bilingual dictionary is available. In this paper we propose a novel approach for query translation disambiguation, named “**spectral query translation model**”. The proposed approach views the problem of query translation disambiguation as a graph partitioning problem. For a given query, a weighted graph is first created for all possible translations of query words based on the co-occurrence statistics of the translation words. The best translation of the query is then determined by the most strongly connected component within the graph. The proposed approach distinguishes from previous approaches in that the translations of all query words are estimated simultaneously. Furthermore, translation probabilities are introduced in the proposed approach to capture the uncertainty in translating queries. Empirical studies with TREC datasets have shown that the spectral query translation model achieves a relative 20% - 50% improvement in cross-language information retrieval, compared to other approaches that also exploit word co-occurrence statistics for query translation disambiguation.

Introduction

Query translation has been an effective way to bridge the gap between the source and target languages in cross-language information retrieval (CLIR). To translate queries from the source language into the target language, it requires external linguistic resources, among which parallel corpora or bilingual dictionaries are the most commonly used. Methods based on parallel corpora, such as relevance language models (Lavrenko, Choquette, & Croft 2002) and statistical translation models (Kraaij, Nie, & Simard 2003; Xu & Weischedel 2001), usually learn an association between words in the source language and the target language, and apply the association to estimate translations of queries. The main drawback of these methods is that they depend critically on the availability of parallel bilingual corpora, which are often difficult to acquire, especially for minor languages. Thus, dictionary-based approaches are usually more

preferable because of the easy access to bilingual dictionaries. However, compared to the corpus-based approaches, dictionary-based approaches usually lack the ability in disambiguating query translations. A simple dictionary-based approach forms the translation of a query by including all the translations of query words provided by the dictionary (which we call *translation candidates* in this paper). But some of the translation candidates could be irrelevant to the original query. Thus, a successful dictionary-based approach should be able to resolve the translation ambiguity in its best effort, and meanwhile preserve the uncertainty in query translation when the ambiguity is hard to reduce.

In the past, several approaches (Adriani 2000a; Gao *et al.* 2001; 2002; Jang, Myaeng, & Park 1999; Kraaij & Pohlmann 2001) have been proposed to resolve the query ambiguity. Given a query in the source language, for each translation candidate of a query word, a coherence score is computed based on its similarity to the query. The translation candidates with the highest coherence scores are selected to form the final translation of the original query. We refer to these approaches as *selection-based approaches*. One of the main problems with the selection-based approaches is that the translation(s) of one query word is usually determined independently from the translations of other query words, which we call “*translation independence assumption*”. Another problem with a selection-based approach is that a *binary* decision is made with regard to whether a translation candidate will be included in the translated query. Given the short length of queries and the large variance existed in mapping information across different languages, such binary decisions are usually difficult, if not impossible, to make. We call this problem the “*translation uncertainty problem*”.

To address the above problems, we propose a novel approach for dictionary-based CLIR, named “**spectral query translation model**”. It views the query translation disambiguation from the perspective of graph partitioning. For a given query, an undirected and weighted graph is constructed for the set of translation candidates of query words: each vertex in the graph corresponds to a unique translation candidate; for any two translation candidates related to two different query words, a weight proportional to their similarity is assigned to the edge connecting them. The best translation of the query corresponds to the most strongly connected

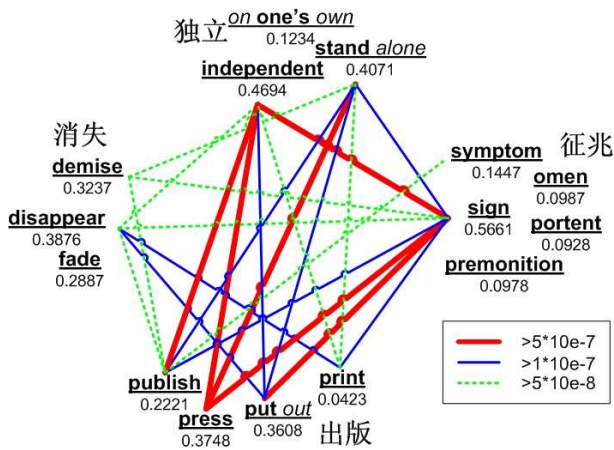


Figure 1: An example of graph partitioning perspective for query translation disambiguation.

clique within the graph, which can be efficiently identified by a spectral clustering method. Since the translation of all query words are determined simultaneously, we are able to avoid the translation independence assumption. To account for the uncertainty in query translation, “soft memberships” will be used in the clustering procedure through the introduction of translation probabilities. An example of applying graph partitioning methods to disambiguate query translations is shown in Figure 1. The query is composed of four Chinese words, and around each Chinese word are its translation candidates in English provided by a Chinese-English dictionary. The thickness of lines connecting two English words roughly represents their correlation. The number below each English word is its translation probability estimated from the proposed approach. Based on the graph representation in Figure 1, we can easily see that the strongly connected component consists of words “independent”, “sign”, and “press”, which have been assigned with large translation probabilities.

The rest of the paper is structured as follows: after briefly reviewing the related work in query translation disambiguation and spectral clustering, we describe our spectral query translation model and the procedures for solving the related optimization problem. Experimental results with analysis will be provided by the end of this paper before we conclude this work.

Related Work

Selection-based CLIR Query Translation

The simplest approach in dictionary-based CLIR is to use all the translation candidates for every query word equally (Davis 1996; Kraaij & Pohlmann 2001). This amounts to no sense disambiguation for query words. Other approaches try to resolve the translation ambiguity by selecting a subset of the translation candidates that are provided by the dictionary. Ideally, for each word in a query we should select the translation(s) that is coherent with the selected translations for other query words. In other words, the selection of trans-

lations for one query word should depend on the translations for other query words. However, due to the computational concern, most selection-based approaches (Adriani 2000a; Gao *et al.* 2001; 2002) adopted an approximate solution: for any translation candidate of a query word, its similarities to all the translation candidates for other query words are computed and summed as its *coherence score*; then, for each query word, the translation candidate with the highest score is selected as the final translation for the query word. Given that the coherence score of each translation candidate is computed based on both selected and unselected translation candidates, this approximation leads to the translation independence assumption that has been discussed in the previous section. In addition to selecting the most likely translation candidate for each query word (Gao *et al.* 2001; Adriani 2000b; Kraaij & Pohlmann 2001), other selection-based approaches have been studied, including selecting the best N translation candidates (Davis 1996) and selecting translations by a predefined threshold (Jang, Myaeng, & Park 1999; Maeda *et al.* 2000).

Spectral Clustering

Spectral clustering approaches view the problem of data clustering as a problem of graph partitioning. Each data point corresponds to a vertex in the graph. Any two data points are connected by an edge whose weight is the similarity between the two data points. To form data clusters, the graph is partitioned into multiple disjoint sets such that only the edges with small weights are removed. Based on different criteria imposed on the partitioning, there are three major variants for spectral clustering: Ratio Cut (Chung 1997), Normalized Cut (Shi & Malik 2000) and Min-Max Cut (Ding *et al.* 2001). In the following, we briefly recapitulate the 2-way Normalized Cut algorithm.

Let $G(V, E, \mathbf{W})$ denote an undirect graph, where V is the vertex set, E is the edge set, and $\mathbf{W} = (w_{i,j})_{n \times n}$ is a matrix with $w_{i,j} \geq 0$ denoting the edge weight between the i -th and the j -th vertex. Define $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_n)$, where $d_i = \sum_{j \in V} w_{i,j}$. To partition the vertex set into two disjoint sets A and B , a 2-way Normalized Cut algorithm minimizes the following objective function:

$$J = \frac{S(A, B)}{d_A} + \frac{S(A, B)}{d_B} \quad (1)$$

where we define $S(A, B) = \sum_{i \in A} \sum_{j \in B} w_{i,j}$ and $d_A = \sum_{i \in A} d_i$. By relaxing cluster memberships to real values, the Normalize Cut algorithm can be formulated into the following eigenvector problem:

$$(\mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{W} \mathbf{D}^{-\frac{1}{2}}) \tilde{\mathbf{q}} = \lambda \tilde{\mathbf{q}} \quad (2)$$

where vector $\tilde{\mathbf{q}}$ is related to the cluster memberships.

Spectral Query Translation Model

The essential idea of the spectral query translation model is to transform the query translation disambiguation problem into a 2-way graph partitioning problem. In the following subsections, we will first describe the graph partitioning view to the query disambiguation problem. Then, we

will show how to introduce translation probabilities into the graph partitioning algorithm, followed by the procedure for solving the related optimization problem. At last, we will show a retrieval model that utilizes the estimated translation probabilities.

The following terminology and notations will be used throughout the rest of the paper. The term ‘‘source language’’ and a superscript s are used when referring to the language of queries. Similarly, the term ‘‘target language’’ and a superscript t are for the language of documents. Let a query of the source language be denoted by $\mathbf{q}^s = \{w_1^s, w_2^s, \dots, w_{m^s}^s\}$, where m^s is the number of distinct words in \mathbf{q}^s . Let \mathbf{r}_k denote the set of translation candidates provided by the dictionary for a word w_k^s in the query \mathbf{q}^s . The whole translation candidate set for the entire query \mathbf{q}^s is then denoted by $\mathbf{R} = \bigcup_{k=1}^{m^s} \mathbf{r}_k$. And we use m^t to denote the size of \mathbf{R} , i.e., the total number of distinct translation candidates.

Query Translation Disambiguation Through Graph Partitioning

Our graph partitioning view of query translation disambiguation can be formally described as follows.

For a given query \mathbf{q}^s and its translation candidate set \mathbf{R} , an undirected weighted graph is created. Each translation candidate $w_k^t \in \mathbf{R}$ is represented by a vertex. Any two translation candidates related to two different query words are connected by an edge if they ever co-occur in at least one document. A non-negative weight is assigned to each edge to indicate the similarity between the two connected words. Among many different co-occurrence statistics, we adopted a variant of mutual information as the similarity measurement, which has been used in previous studies (Gao *et al.* 2002)

$$s_{j,j'}^t = \Pr(w_j^t, w_{j'}^t) \times \log \frac{\Pr(w_j^t, w_{j'}^t)}{\Pr(w_j^t) \times \Pr(w_{j'}^t)} \quad (3)$$

$\Pr(w_j^t)$ is the unigram probability for word w_j^t , and $\Pr(w_j^t, w_{j'}^t)$ is the joint probability for word w_j^t and $w_{j'}^t$ to co-occur in the same documents. Note that Equation (3) is different from the standard definition for mutual information in that only co-occurrence information is used. Due to the computation concern, in Equation (3) we ignore the correlation between two words when at least one of them does not occur in documents.

With the constructed graph for a given query, we hypothesize that the best translation of a query corresponds to the most strongly connected component within the graph. To separate the strongly connected component from the rest of the graph, a graph partitioning algorithm can be employed to divide the graph into two disjoint clusters: a cluster for strongly connected component, and a cluster for the rest of the graph. To this end, we inherit the idea from the Normalized Cut algorithm. For the graph constructed for the translation candidate set \mathbf{R} , we define the adjacency matrix as $\mathbf{S} = [s_{j,j'}^t]_{m^t \times m^t}$ where $s_{j,j'}^t$ is the similarity measurement defined in Equation (3). Let diagonal matrix $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_n)$ where $d_j = \sum_{j'=1}^{m^t} s_{j,j'}^t$. Then

the graph Laplacian matrix is $\mathbf{L} = \mathbf{D} - \mathbf{S}$. Following the formalism of the Normalized Cut algorithm, the optimal 2-way partitioning is found by minimizing the following objective function:

$$J = \mathbf{v}^T \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}} \mathbf{v} \quad (4)$$

Here $\mathbf{v} = [v_1 v_2 \dots v_{m^t}]^T$ is a cluster indicator vector. Each element v_i is a binary variable with 1 indicating the corresponding word being included in the query translation and 0 for not being included. For later reference, we name this model ‘‘spectral query translation model’’, or ‘‘SQT’’ for short.

Soft Cluster Memberships via the Introduction of Translation Probabilities

To address the ‘‘translation uncertainty problem’’ mentioned in the introduction section, we introduce translation probability in our new model.

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

$$p_{k,j} = \Pr(w_j^t | w_k^s, \mathbf{q}^s) \quad (5)$$

which satisfies

$$\begin{aligned} &\text{if } w_j^t \notin \mathbf{r}_k, p_{k,j} = 0; \text{ otherwise } p_{k,j} \geq 0 \\ &\sum_{\forall j, w_j^t \in \mathbf{r}_k} p_{k,j} = 1 \end{aligned}$$

To simplify our notation, matrix $\mathbf{P} = [p_{k,j}]_{m^s \times m^t}$ is used to denote all the translation probabilities associated with the query \mathbf{q}^s . Let $\mathbf{T} = [t_{k,j}]_{m^s \times m^t}$ denote the part of the bilingual dictionary that is related to query \mathbf{q}^s . An element $t_{k,j}$ in \mathbf{T} is 1 if the word w_j^t appears as a translation in the dictionary for the word w_k^s , and 0 otherwise. Then the above constraints can be rewritten as

$$\mathbf{P} \cdot \mathbf{1}_{m^t \times m^t} = \mathbf{I} \quad (6)$$

$$\mathbf{0} \leq \mathbf{P} \leq \mathbf{T} \quad (7)$$

where $\mathbf{1}_{m^t \times m^t}$ is a matrix of all elements of 1’s and $\mathbf{I} = \text{diag}(1, 1, \dots, 1)$.

Furthermore, we compute the probability for a translation candidate to be adopted in the query translation, i.e.,

$$\Pr(w_j^t | \mathbf{q}^s) = \sum_{w_k^s \in \mathbf{q}^s} \Pr(w_j^t | w_k^s, \mathbf{q}^s) \Pr(w_k^s | \mathbf{q}^s) \quad (8)$$

Here $\Pr(w_k^s | \mathbf{q}^s)$ is from a monolingual language model for query \mathbf{q}^s in the source language. For the sake of simplicity, we assume a uniform language model for the query \mathbf{q}^s , i.e., $\Pr(w_j^t | \mathbf{q}^s) = p_{k,j} / m^s$.

From the graph partitioning perspective, the probability $\Pr(w_j^t | \mathbf{q}^s)$ can be interpreted as a soft membership of the translation candidate w_j^t to the most strongly connected cluster. Instead of constraining each element v_i in the cluster indicator vector \mathbf{v} to be a binary variable, we can set $v_i = \Pr(w_i^t | \mathbf{q}^s)$. Then, each element v_i indicates how likely the corresponding translation candidate will be included in

the strongly connected component, or the final translation for the query \mathbf{q}^s .

Finally, combining Equation (5)-(8), the query translation disambiguation can be formulated into the following optimization problem:

$$\begin{aligned} \min_{\mathbf{P}} \mathbf{e}^T \mathbf{P} \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}} \mathbf{P}^T \mathbf{e} \quad (9) \\ \text{s.t. } \mathbf{P} \cdot \mathbf{1}_{m^t \times m^t} = \mathbf{I} \\ \mathbf{0} \leq \mathbf{P} \leq \mathbf{T} \end{aligned}$$

where \mathbf{e} is a vector with all elements as 1s. This is a convex programming problem, where the unique optimal solution is guaranteed. Note that, by solving Equation (9), we are able to estimate the translation probabilities for *all* query words simultaneously, thus removing the ‘‘translation independence assumption’’.

Solving the optimization problem

The optimization problem in (9) is in fact a standard quadratic programming (QP) problem (Gill, Murray, & Wright 1981). To give it an explicit QP form, we define

$$\mathbf{P}_{m^s \times m^t} = \begin{pmatrix} \mathbf{p}_1^T \\ \mathbf{p}_2^T \\ \vdots \\ \mathbf{p}_{m^s}^T \end{pmatrix} \quad \mathbf{T}_{m^s \times m^t} = \begin{pmatrix} \mathbf{t}_1^T \\ \mathbf{t}_2^T \\ \vdots \\ \mathbf{t}_{m^s}^T \end{pmatrix} \quad (10)$$

$$\tilde{\mathbf{P}}_{m^s m^t \times 1} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_{m^s} \end{pmatrix} \quad \bar{\mathbf{P}}_{m^s m^t \times 1} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \vdots \\ \mathbf{t}_{m^s} \end{pmatrix} \quad (11)$$

$$\mathbf{A}_{m^s m^t \times m^s m^t} = \mathbf{1}_{m^s \times m^s} \otimes (\mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}}) \quad (12)$$

$$\mathbf{E}_{m^s \times m^s m^t} = \text{diag}(\mathbf{t}_1^T, \mathbf{t}_2^T, \dots, \mathbf{t}_{m^s}^T) \quad (13)$$

where \otimes represents *kroncker product*. Then, the optimization problem in (9) can be rewritten in a standard form of the QP problem as follows:

$$\begin{aligned} \max_{\tilde{\mathbf{p}}} \tilde{\mathbf{p}}^T \mathbf{A} \tilde{\mathbf{p}} \quad (14) \\ \text{s.t. } \mathbf{E} \tilde{\mathbf{p}} = \mathbf{1}_{m^s \times 1} \text{ and } 0 \leq \tilde{\mathbf{p}} \leq \bar{\mathbf{p}} \end{aligned}$$

In our experiment, the QP package from MATLAB is used to solve the above problem.

Retrieval Model

The introduction of translation probabilities $p_{k,j}$ can be well accommodated by a statistical retrieval model for CLIR. 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, the logarithm of this probability can be approximated as

$$\begin{aligned} \log \Pr(\mathbf{d}^t | \mathbf{q}^s) &\sim \log \Pr(\mathbf{q}^s | \mathbf{d}^t) \\ &\sim \sum_{w^t} \Pr(w^t | \mathbf{q}^s) \log \Pr(w^t | \mathbf{d}^t) \quad (15) \end{aligned}$$

Here $\Pr(w^t | \mathbf{d}^t)$ is a monolingual language model for document \mathbf{d}^t in the target language; $\Pr(w^t | \mathbf{q}^s)$ is the soft membership for translation candidate w^t to be in the most strongly connected cluster and can be computed from the set of translation probabilities as in Equation (8).

Experiments

Our experiments are designed to examine the effectiveness of the proposed model for cross-language information retrieval in the following two aspects:

1. Is the proposed spectral query translation model effective for CLIR?
2. How important is the translation uncertainty and the removal of the translation independence assumption for CLIR?

The first aspect is examined by comparing the proposed method to existing approaches, and the second aspect is addressed through case studies.

Experiment Setup

All our experiments are retrieval of English documents using Chinese queries. TREC ad hoc test collections are used in our experiments, 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 abstracts¹

In addition to the homogeneous collections listed above, we also tested the proposed model against two heterogeneous collections that are formed by combining multiple homogeneous collections: collection AP88-89 + WSJ87-88, and collection AP89 + WSJ87-88 + DOE1-2. In a heterogeneous collection, words are more likely to carry multiple senses, thus increasing the difficulty in word sense disambiguation. The SMART system (Salton 1971) is used to process document collections with stop words removal and word stemmed.

Our queries come from a manual Chinese translation of TREC-3 ad hoc topics (topic 151-200). Both short and long Chinese queries are tested: short ones are created by translating the ‘‘title’’ field of English queries into Chinese and long ones are formed by combining the Chinese translations of both the ‘‘title’’ and ‘‘description’’ fields in English queries. 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. The Chinese-English dictionary from Linguistic Data Consortium (LDC, <http://www ldc upenn edu>) is used in our experiment. Since our experiments do not involve the processing of English phrases, any English phrase in the translation of a Chinese word is treated as a bag of words.

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

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 spectral query translation model over the other two methods)

	SHORT QUERIES					LONG QUERIES				
	BSTO	ALTR	SQT	(S-B)/B	(S-A)/A	BSTO	ALTR	SQT	(S-B)/B	(S-A)/A
AP	0.2381	0.2241	0.3116	+30.87%	+39.05%	0.1749	0.1803	0.2426	+38.71%	+34.55%
WSJ	0.1966	0.2129	0.2571	+30.77%	+20.76%	0.1478	0.1727	0.2161	+46.21%	+25.13%
AP+WSJ	0.2253	0.2209	0.2859	+26.90%	+29.43%	0.1433	0.1665	0.2048	+49.92%	+23.00%
AP+WSJ+DOE	0.1739	0.1829	0.2296	+32.03%	+25.53%	0.1122	0.1411	0.1712	+52.58%	+21.33%

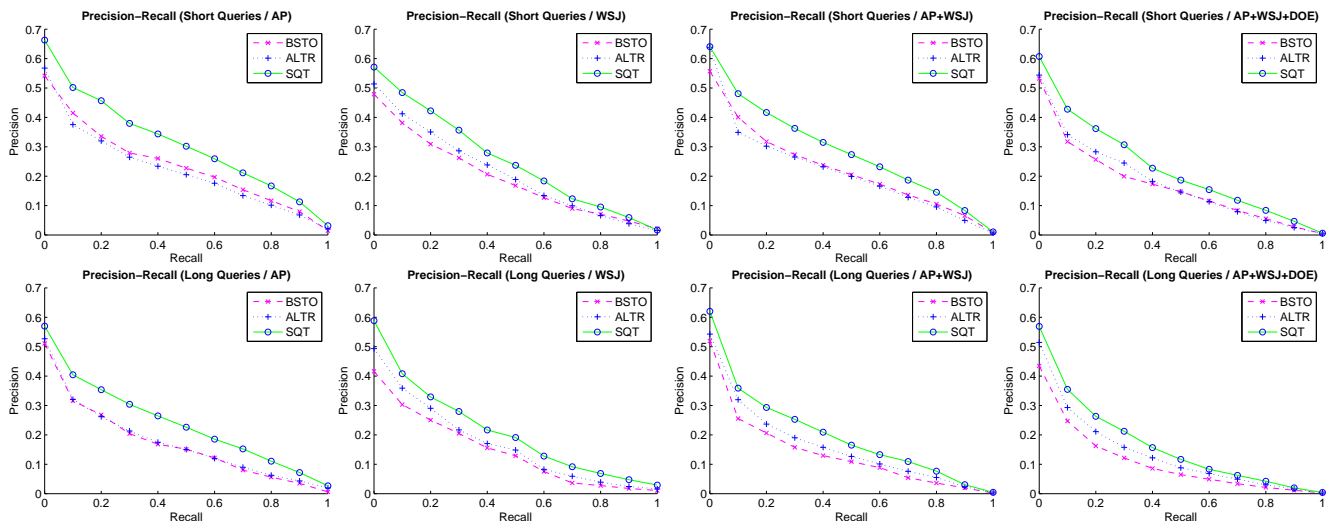


Figure 2: Comparison of CLIR performance on four datasets using both short and long queries. The first two columns are for homogeneous datasets: AP88-89 and WSJ87-88; the last two columns are for heterogeneous datasets: AP88-89 + WSJ87-88 and AP89 + WSJ87-88 + DOE1-2. The upper row is for short queries, and the lower row is for long queries.

Two selection-based approaches are used as the baseline models. The first one selects the most likely translation candidate for each query word based on the algorithm in (Adriani 2000a), which we call “BSTO”. The second model, which we call “ALTR”, includes all the translation candidates for query words into the final query translation.

Comparison with Selection-based Approaches

Table 1 lists the average precision across 11 recall points for both the homogeneous and heterogeneous collections. As indicated in Table 1 the proposed model (i.e., “SQT”) is able to outperform the two baseline models for both short and long queries across all the four different collections. Furthermore, we plot the precision-recall curves for both the short queries and the long queries in Figure 2, respectively. We clearly see that for all the four collections, the precision-recall curves of the spectral query translation model stay above the curves of the other two models. Based on these results, we conclude that the spectral query translation 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 is worse than that for homogeneous collections.

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 a higher translation ambiguity.

2. A better retrieval performance is achieved for short queries than for long queries. Furthermore, the performance gap between the short queries and the long 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, particularly when the disambiguation algorithm is based on word similarities.
3. The “BSTO” method does not consistently outperform the “ALTR” method. In fact, for the long queries, the “ALTR” method performs better than the “BSTO” method across all four 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. Thus, 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 (Gao *et al.* 2001). However, the setup of our experiments is rather different from theirs. For example, we did

not identify English phrases in our text processing, which has been shown to be important for CLIR (Ballesteros & Croft 1997). Despite the importance of phrase analysis for CLIR, we believe that a generic probabilistic model will be beneficial to CLIR of any languages, particularly when linguistic resources are scarce.

The Impact of Translation Uncertainty and Translation Independence Assumption on CLIR

To demonstrate the uncertainty in query translation, we reuse the example in Figure 1, where each translation candidate is annotated with its translation probability. A variance can be observed in the distribution of translation probabilities across the four Chinese words: for the word on the left, its translation probability distribution is close to be uniform; for the word on the right, its distribution is rather skewed; for the rest two words, their translation probability distributions are between the two extremes. This variance suggests the difficulty to apply the selection-based approaches for query translation disambiguation, since they only make binary decision in selecting translations.

The example in Figure 1 also reveals the impact of translation independence assumption on query translation disambiguation. In this example, the translations selected by the “BSTO” method are “sign”, “press”, “disappear” and “stand”². Given the context of the original query, “independent” should be a better translation than “stand”. One reason for such a mistake is that in the “BSTO” method, the coherence score of a translation is computed based on all the English words in the translation candidate set. Since “stand” is common in English, it co-occurs with many other English words. Although the mutual information for each co-occurring word is small, the overall coherence score for “stand” turns out to be larger than that of “independent”. In contrast, spectral query translation model eliminates this problem by simultaneously estimating the translation probabilities of all translation candidates.

Conclusions

In this paper, we propose a novel method, named “spectral query translation model”, that applies a graph partitioning algorithm to disambiguate query translation in CLIR. Compared to the selection-based approaches, our model is advantageous in two aspects: 1) It maintains the translation uncertainty through the estimation of translation probabilities; 2) The simultaneous estimation of all translations allows the proposed model to avoid the translation independence assumption, hence forming more coherent query translations. Empirical studies with CLIR under various scenarios have shown that the proposed model is able to perform substantially better for CLIR than several existing selection-based approaches.

References

Adriani, M. 2000a. Dictionary-based CLIR for the CLEF multilingual track. In *CLEF '00*.

²The word “stand” corresponds to the phrase “stand alone”. “Alone” is removed because it is a stop word.

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

Ballesteros, L., and Croft, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *SIGIR '97*, 84–91. ACM Press.

Chung, F. R. K. 1997. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, ISSN: 0160-7642. American Mathematical Society.

Davis, M. W. 1996. New experiments in cross-language text retrieval at NMSU’s computing research lab. In Harman, D. K., ed., *TREC-5*. NIST.

Ding, C. H. Q.; He, X.; Zha, H.; Gu, M.; and Simon, H. D. 2001. A min-max cut algorithm for graph partitioning and data clustering. In *ICDM '01*, 107–114. IEEE Computer Society.

Gao, J.; Nie, J.-Y.; Xun, E.; Zhang, J.; Zhou, M.; and Huang, C. 2001. Improving query translation for cross-language information retrieval using statistical models. In *SIGIR '01*, 96–104. ACM Press.

Gao, J.; Zhou, M.; Nie, J.-Y.; He, H.; and Chen, W. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *SIGIR '02*, 183–190. ACM Press.

Gill, P.; Murray, W.; and Wright, M. 1981. *Practical Optimization*. San Diego, USA: Academic Press, Inc.

Jang, M.-G.; Myaeng, S. H.; and Park, S. Y. 1999. Using mutual information to resolve query translation ambiguities and query term weighting. In *ACL '99*.

Kraaij, W., and Pohlmann, R. 2001. Different approaches to cross language information retrieval. In Daelemans, W.; Sima’an, K.; Veenstra, J.; and Zavrel, J., eds., *Computational Linguistics in the Netherlands 2000*, number 37 in Language and Computers: Studies in Practical Linguistics, 97–111. Amsterdam: Rodopi.

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

Lavrenko, V.; Choquette, M.; and Croft, W. B. 2002. Cross-lingual relevance models. In *SIGIR '02*, 175–182. ACM Press.

Maeda, A.; Sadat, F.; Yoshikawa, M.; and Uemura, S. 2000. Query term disambiguation for web cross-language information retrieval using a search engine. In *IRAL '00*, 25–32. ACM Press.

Salton, G., ed. 1971. *The SMART retrieval system*. Prentice-Hall.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. on PAMI* 22(8):888–905.

Xu, J., and Weischedel, R. 2001. TREC-9 cross-lingual retrieval at BBN. In *TREC-9*.