Combining Lexical and Statistical Translation Evidence for Cross-Language Information Retrieval

Sungho Kim[+], Youngjoong Ko[+] and Douglas W. Oard[*]

[+]Computer Engineering, Dong-A University, Busan, 604-714 Korea
*College of Information Studies & UMIACS, University of Maryland, College Park, MD 20742 USA
knife16@gmail.com, youngjoong.ko@gmail.com, oard@umd.edu

1. ***Corresponding author :*** Youngjoong Ko
2. ***Corresponding address :*** Department of Computer Engineering, Dong-A University, 840, Hadan 2-dong, Saha-gu, Busan, 604-714, Korea
3. ***Corresponding telephone number :*** 82-51-200-7782
4. ***Corresponding fax number :*** 82-51-200-7783
5. ***Corresponding Email address :*** youngjoong.ko@gmail.com

Abstract

This paper explores how best to use lexical and statistical translation evidence together for Cross-Language Information Retrieval (CLIR). Lexical translation evidence is assembled from Wikipedia and from a large machine readable dictionary, statistical translation evidence is drawn from parallel corpora, and evidence from co-occurrence in the document language provides a basis for limiting the adverse effect of translation ambiguity. Coverage statistics for NTCIR queries confirm that these resources have complementary strengths. Experiments with translation evidence from a small parallel corpus indicates that even rather rough estimates of translation probabilities can yield further improvements over a strong technique for translation weighting based on using Jensen-Shannon divergence as a term association measure. Finally, a novel approach to post-translation query expansion using a random walk over the Wikipedia concept link graph is shown to yield further improvements over alternative techniques for post-translation query expansion. Evaluation results on the NTCIR-5 English-Korean test collection show statistically significant improvements over strong baselines.

Key Words: Cross-Language Information Retrieval, Wikipedia, Small Parallel Corpus

1. Introduction

In Cross-Language Information Retrieval (CLIR), the objective is to find (and usually to rank) documents that are expressed in one language based on queries that are expressed in another (Oard, 2009). Performing CLIR requires some way of mapping terms from one language to another. Research on CLIR has to date for the most part been balkanized into two loosely connected literatures, each focused principally on the source of evidence for those translation[1] mappings. One broad class of techniques, which were developed for the most part in the decade of the 1990's, but which continue to receive attention today, was initially known as "dictionary-based CLIR," although in this paper we refer to it more generally as being based on lexical evidence (i.e., evidence assembled from some hand-built lexicon or lexicons) (Pirkola, 1998). An alternative approach, also developed initially in the 1990's but really coming to the fore in the decade of the 2000's, relies instead on the observed statistics of language use for evidence of the meaning similarity. Initially, such statistics were obtained from comparable corpora (separately authored texts with similar meanings), but the rapid adoption of statistical alignment of translated text as a basis for machine translation in the late 1990's and early 2000's led to statistical evidence from automated alignment of parallel (i.e., translation-equivalent) corpora garnering much of the research attention among what we refer to broadly in this paper as "statistical" techniques (Landauer & Littman, 1990; McCarley, 1999). Machine Translation (MT) systems could be used to implement either approach – one-best rule-based MT as one way of implementing dictionary-based CLIR, one-best statistical MT as one way of implementing CIR based on parallel text. These literatures have not been completely disjoint, of course. For example, the coupled use of pre-translation and post-translation query expansion can be seen as a way of using lexical evidence of translation equivalence to exploit unlinked comparable corpora (i.e., separately authored texts for which topical relationships between specific document pairs are not known *a priori*) (Ballesteros & Croft, 1997). Nonetheless, to date there has been far less attention paid to how to use lexical and statistical translation evidence together than there has been to how to use them separately.

For this paper, we focus on using lexical and statistical evidence together. We believe that this is a natural question to explore because the two types of evidence exhibit complementary strengths.

[1] In keeping with common usage, we refer to cross-language term mappings as "translations," but what is really meant is that the meaning to the terms is related in a way that is useful for information retrieval. Many types of relationships might be useful, including equivalence, hyponymy, and contextual co-occurrence (e.g., doctor and nurse have different meanings, but they co-occur in similar contexts).

Statistical evidence (when drawn from parallel text) can be nuanced (in the sense that alignment probabilities can tell us which term translations are most likely to have been used), but modeling rare translations using statistical evidence alone can be problematic (unless we have access to very large domain-matched parallel corpora). This can be an important limitation because users of information retrieval systems will naturally sometimes craft queries using relatively rare terms in an effort to be precise. For lexical evidence, comparison over a broad range of Web accessible bilingual dictionaries has shown that the coverage (in TREC queries) of query terms other than proper names becomes quite good once dictionaries have a few hundred thousand terms (Demner-Fushman & Oard, 2003). Proper names are, however, often present in queries. Recently, CLIR research has turned to Wikipedia as a source of lexical (or statistical) translation evidence, in part because Wikipedia has good coverage of many proper names that might occur in queries (Giallard et al, 2010; Sorg & Cimiano, 2008). On the other hand, Wikipedia's translation mappings are more limited than Wikipedia's coverage in any one language, so evidence from Machine-Readable Dictionaries (MRD) can still be useful. However, a key limitation of any rich manually created source of lexical evidence is substantial overgeneration of translation alternatives, for the simple reason that language is inherently ambiguous. This has led to a small cottage industry of techniques for leveraging co-occurrence evidence for translation selection that statistical machine translation researchers would recognize as variations on language modeling (Federico & Bertoldi, 2002).

In this paper, we make four principal contributions. First, we use Wikipedia together with a large MRD in order to couple the balanced translation coverage of our MRD with the richer proper name translation coverage of Wikipedia. Second, we combine this translation evidence from lexical resources with statistical translation evidence from a relatively small parallel text collection, using a unified language modeling approach to weight the translation alternatives. Third, we introduce a novel approach to post-translation query expansion, using evidence assembled from the link structure of Wikipedia. Fourth, using these techniques together, we demonstrate substantial and statistically significant improvements over the best previously reported results for CLIR with English queries and Korean documents on a standard NTCIR test collection. Korean is an example of a language for which large-scale lexical resources are available, but for which very large parallel text collections have yet to be assembled. It thus provides an appropriate test case for the utility of the techniques that we present.

The remainder of this paper is organized as follows. Section 2 presents related work on CLIR using MRD, Wikipedia, and parallel text. Section 3 then introduces the linguistic resources that we used as a basis for query translation and query expansion, Section 4 explains how we used them together, and Section 5 describes some baseline techniques to which we compare. Section 6 details our experiment design, and Section 7 presents our results. Section 8 concludes the paper with some remarks on future work.

2. Related Work

The defining characteristic of CLIR is that queries and documents are expressed in different languages (Jones et al., 1999). When the translation mapping is invariant over time, the choice between query translation and document translation is an implementation detail that need not alter the outcome (i.e., the query translation implementations have straightforward document translation implementations in which the translation mapping is simply pre-compiled into the document representation). We therefore focus here, and throughout this paper, on query translation because that approach offers the greatest flexibility for experiment design. For query translation, three basic types of translation resources have been tried: machine translation, corpora, and lexicons (Callan et al., 1992; Robertson and Walker, 1999). Machine translation is, however, simply one way of using translation evidence that comes originally from statistical or lexical sources. We begin therefore with a brief review of related work on CLIR using translation evidence from MRD. We then survey related work on the use of Wikipedia as a source for either translation evidence or query expansion. Finally, we conclude the section with a brief review of the use of statistical translation evidence in CLIR.

2.1. CLIR using Evidence from a MRD

When suitable corpora are not available (and thus when effective machine translation systems are not easily built), bilingual lexicons assembled from a bilingual word list, an MRD, or a bilingual thesaurus can be used as a source for encoded translation relations. Dictionaries typically lack morphological variants, but because the effect of morphological variation is normally suppressed in retrieval through the use of stemming, the simple expedient of stemming both the lexicon and the documents is typically effective (Levow et al, 2005). Because such lexicons typically lack the types of statistical translation preference evidence that can be found using corpora, it can be useful to

instead rely on evidence from the co-occurrence of alternate translations of query terms to constrain the choice of translations. Evidence for this co-occurrence will naturally be available from the document collection to be searched, although when that collection is small it can be useful to draw co-occurrence evidence from other corpora. Seo et al. (2005) have demonstrated one effective way of doing this using rich lexical resources. At the time of that work, such resources were available only for a few language pairs, but today suitable resources are easily assembled from Wikipedia.

2.2. CLIR using Evidence from Wikipedia

Wikipedia has emerged in recent years as a potentially important resource for CLIR generally. One interesting example of the use of Wikipedia in CLIR is the work of Nguyen et al. (2008), who performed query translation using only Wikipedia to obtain translations. Queries were mapped to Wikipedia concepts (i.e., the Wikipedia page on a topic) and the corresponding translations of these concepts in the target language (i.e., linked pages on the same topic) were used to create the final query. Gaillard et al. (2010) investigated two successive steps for translation; finding translation alternatives from Wikipedia cross-lingual links, and disambiguation using Wikipedia categories of target language. Schönhofen et al. (2008) also utilized the linked pairs of English and Hungarian Wikipedia article titles, and then exploited Wikipedia hyperlinks for query term disambiguation. Tang et al. (2010) searched for the best English Wikipedia pages using Google and then followed the inter-wiki links to identify the corresponding Chinese Wikipedia pages, ultimately using the titles of those Chinese pages as their translation of the query.

2.3. Query Expansion using Evidence from Wikipedia

A number of researchers have adopted Wikipedia as a basis for query expansion (Elsas et al., 2008; Fautsch and Savoy, 2008; Mishne, 2007; Zhang and Yu, 2006). Zhang and Yu (2006), for example, combined evidence from different parts of Wikipedia articles for query expansion, and Elsas et al. (2008) notably used cross-article hyperlinks, which we also exploit. Specifically, Elsas et al. used the query to identify related Wikipedia pages, they then identified the anchor text on other Wikipedia pages that linked to those query-related pages, and finally they added terms found in that anchor text to the query as an expansion step.

2.4. CLIR using Evidence from Parallel Corpora

When bilingual corpora are available that use language in ways similar to the language use in the queries and the documents, alternative approaches that do not rely on a machine readable dictionary can be employed. Two types of bilingual corpora can be used for CLIR: parallel corpora, or comparable corpora. A parallel corpus is constructed by actually translating documents between languages. Document alignments result naturally from the process, but sentence and ultimately term alignments must be inferred based on heuristics that capture conventions of the translation process (e.g., consistent sentence ordering in the two languages) and statistical regularities. These alignments can then be used to estimate translation probabilities, which then can be used as a basis for CLIR (Darwish and Oard, 2003). A comparable corpus, by contrast, consists of independently authored documents on related (i.e., "comparable") topics. Because document alignments do not naturally result from the generation process in the case of comparable corpora, they must be inferred in some way. It is possible to draw on statistical regularities to construct a bilingual lexicon from a comparable corpus (McNamee, 2008), but the resulting corpus is rarely as useful as that which could be constructed using parallel text. As a result, the most common way of using comparable corpora for CLIR is to perform query expansion before and after translation using blind relevance feedback techniques (Ballesteros and Crroft, 1997). In this case, the necessary document-scale alignment is performed on the fly as a part of the query (or document) translation process.

3. Linguistic Resources

In this section, we introduce the linguistic resources that we have used in our experiments as a basis for query translation and expansion. Our "Wikipedia-Based Lexicon (WBL)" is based on three lexicons that are automatically extracted from Wikipedia (a bilingual term pair lexicon, two monolingual synonym lexicons, and two monolingual polysemy lexicons) and that are then combined to produce the single pre-compiled translation lexicon that we refer to as WBL in our experiments. In addition, we describe how we construct the Wikipedia concept link graph used in query expansion. Korean is an agglutinative alphabetic language, with spaces between words (as in English) and many compound words (as in German). Modern Korean is written using the Hangul alphabet.[2]

[2] http://en.wikipedia.org/wiki/Hangul

3.1. Creating the Translation Lexicon from Wikipedia

Wikipedia links offer a source of evidence for both cross-language and within-language term relationships. Cross-language mappings are available from the "Inter-wiki" links that link pages on the same topic in different languages. Evidence for within-language synonymy is available from redirect pages, and language-specific polysemy information can be obtained from disambiguation pages. These three lexicons are used together to translate the query; we refer to them as our "Wikipedia-Based Lexicon (WBL)." As an open collaboration, Wikipedia content is readily downloadable.[3] The English and Korean versions used in our experiments were released on May 26, 2011 and June 7, 2011, respectively.

Table 1. WBL: Wikipedia-Based Lexicon

Content	English Wikipedia	Korean Wikipedia
Wikipedia articles	8,389,381	273,606
Bilingual term pairs	105,643	
Synonymy sets	1,034,492	131,213
Polysemy sets	195,390	16,998

1) Constructing the Bilingual Term Pair Lexicon from Inter-Wiki Links

As of January 2012, there were editions of Wikipedia in 283 languages. For example, the English page "President of the United States" has corresponding pages in several languages (French: "Président des États-Unis," Korean: "미국의 대통령," German: "Präsident der Vereinigten Staaten," etc.). These correspondences are expressed in Wikipedia as so-called Inter-wiki links that are included in the body of many articles. Using these links, we constructed a bilingual term pair lexicon. Table 2 provides a few examples.

Table 2. Examples from the bilingual term pair lexicon

English	Korean
Andre Agassi	안드레 애거시
Apache Software Foundation	아파치 소프트웨어 재단
President of South Korea	대한민국의 대통령

2) Constructing English and Korean Synonymy Lexicons from Redirect Pages

[3] http://download.wikipedia.org

Another useful resource found in Wikipedia is the so-called redirect pages, which identify alternative names that can be used to refer to a Wikipedia concept (Pu et al., 2007). For example, the page "U.S.A" redirects to the article "United States" which contains information about that nation. We constructed synonymy lexicons for English and Korean from this information, as illustrated in Table 3. As the example above illustrates, what we call synonymy lexicons also include acronyms and abbreviations. The English and Korean synonymy lexicons operate independently for query expansion before and after translation.

Table 3. Examples from the synonym lexicons

Concept	Synonym	Concept	Synonym
United States	U.S.A	미국	미합중국
United States	USA	미국	아메리카 합중국
United States	UnitedStates	미국	USA

3) Constructing English and Korean Polysemy Lexicons using Disambiguation Pages

Disambiguation pages in Wikipedia are intended to allow users to choose among several Wikipedia concepts for an ambiguous word. In other words, they list the referents of ambiguous words and phrases that denote two or more concepts in Wikipedia. For example, the page "Washington" contains dozens of referents, including "George Washington," "Washington (state)," and "Washington, D.C."

Table 4. Examples from the polysemy lexicons

English Concept	English Senses	Korean Senses
Washington	George Washington	조지워싱턴
Washington	Washington (State)	워싱턴 주
Washington	Washington, D.C.	워싱턴 D.C.

A single pre-compiled translation lexicon (WBL) was assembled from these components by starting with each English term in the bilingual pair lexicon, expanding that term to a set of related English terms by first using the English polysemy lexicon and then the English synonymy lexicon, then translating every word in that set using the bilingual pair lexicon, and then expanding each Korean term in the resulting set by first using the Korean polysemy lexicon and then the Korean synonymy lexicon. All English terms were then stemmed using the Porter stemmer, all Korean terms were

replaced with their morphological root form using the KLT Korean morphological analyzer,[4] and any resulting duplicate term pairs were then removed.

3.2. Bilingual Machine Readable Dictionary

As an encyclopedia, Wikipedia is a better source for translations of highly specific terms than of common terms. We therefore also obtained an English-to-Korean Machine Readable Dictionary (MRD),[5] which (before stemming) contains 200,195 unique English terms, 885,642 unique Korean terms, and 1,851,587 English-Korean translation pairs. For our MRD Lexicon, all English terms are stemmed, all Korean terms are replaced with their morphological root form, and duplicate term pairs are then removed. For exact term match for translation candidate generation in Section 4.1, we also maintain the original terms before stemming and morphological analysis. The term matching process is conducted by two steps. The first step attempts to match an unstemmed query term to any unstemmed query term in the lexicon. If there is no match, it does the stemmed query term is matched to any stemmed lexicon term.

3.3. Combined Translation Lexicon

We created a Combined Translation Lexicon (CTL) by taking all Korean translations for every English term that was present in the WBL lexicon from that lexicon and, for other English terms that were found in the MRD, by taking all Korean translations from that MRD. Table 5 shows the coverage for the MRD Lexicon, the WBL, and the CTL. The values are expressed as the number of matching terms from the 50 English NTCIR-5 title queries that we used in the experiments described below in Section 6, after stopword removal.

Table 5. Matching English Title query terms

Title	MRD	WBL	CTL
Queries	50		
Unique query terms	130		
Total query terms	240		
Matched query terms	150	163	217
Coverage	62%	68%	90%
Total translations	640	250	559
Avg. trans./query term	4.26	1.53	2.58

[4] http://nlp.kookmin.ac.kr/HAM/kor/
[5] http://dic.daum.net

As can be seen in Table 5, the ratio of matched English terms if we used the MRD Lexicon alone would be 62%, with an average of 4.26 translations per English term. The WBL matches somewhat more of the English terms (68%), and moreover it does so with considerably less translation ambiguity (averaging only 1.53 Korean translations per English term). Notably, the CTL matches far more terms than either of the two lexicons from which it is built (90%), with an average number of translations that falls between that of the two other lexicons (2.58). From this, we can conclude that the WBL and MRD Lexicons have largely complementary coverage.

Multi-word expressions are typically less ambiguous than single words, so bilingual lexicons with a substantial number of multi-word expressions on the source side can help to limit translation fanout. As Table 6 shows, Wikipedia provides an excellent source for such translation pairings, with well over half of all translation pairs containing a multi-word expression on the English side. Some multi-word expressions happen to be compositional; others are idiomatic expressions. We do not distinguish between the two cases. Our machine readable dictionary is also richer in English multi-word expressions than would be expected from an English-to-Korean bilingual dictionary, reflecting its origin as a Korean-to-English bilingual dictionary for which we are using the translation mappings in the reverse direction.

Table 6. Number of English multi-word expressions in the WBL and MRD lexicons

	WBL	**MRD**
Total terms	105,643	200,195
Multi-word expressions	62,942	23,895
Multi-word expression fraction	59.6 %	11.9 %

3.4. Parallel Text

To learn statistical translation mappings, we use the sentence-aligned 21st Century Sejong Project Korean-English Parallel Corpus,[6] which contains 52,998 sentence pairs. We tokenized and stemmed the English, converted the Korean to root forms, and aligned the resulting English stems and Korean roots using Giza++.[7] We then computed unsmoothed translation probabilities from the resulting alignment counts (see Section 4 below for our experiments with smoothing).

[6] http://www.sejong.or.kr
[7] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

3.5. Wikipedia Concept Link Graph for Query Expansion

We expect the cross-article hyperlinks between same-language Wikipedia pages to provide useful evidence for relationships between concepts. This seems like a reasonable assumption because the body of each article describes a concept, and links to other Wikipedia pages (i.e., to other concepts) will be included within the page when the author believes that those other concepts are useful as a part of that description. Therefore, based on the cross-article hyperlinks found in Wikipedia, we construct a directed concept link graph $G=(X,E)$, where X is a set of Wikipedia articles that is represented as $X = \{x_i\}_{i=1}^{m}$, m is the total count of Wikipedia articles, and each edge in E connects two articles. Let W represent the $m \times m$ weight matrix of graph G, in which element w_{ij} equals the link count associating vertices between x_i and x_j in the matrix. We use this concept link graph as a basis for query expansion in Section 4.

4. Proposed Technique

In the section, we describe our query translation and expansion methods. We begin with an overview of the query translation process (see Figure 1) and then describe each step in detail.

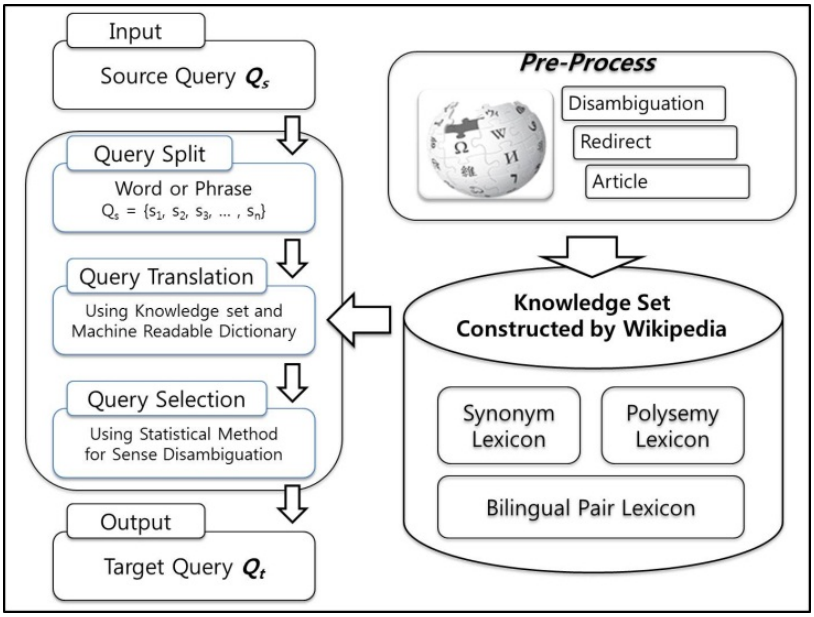
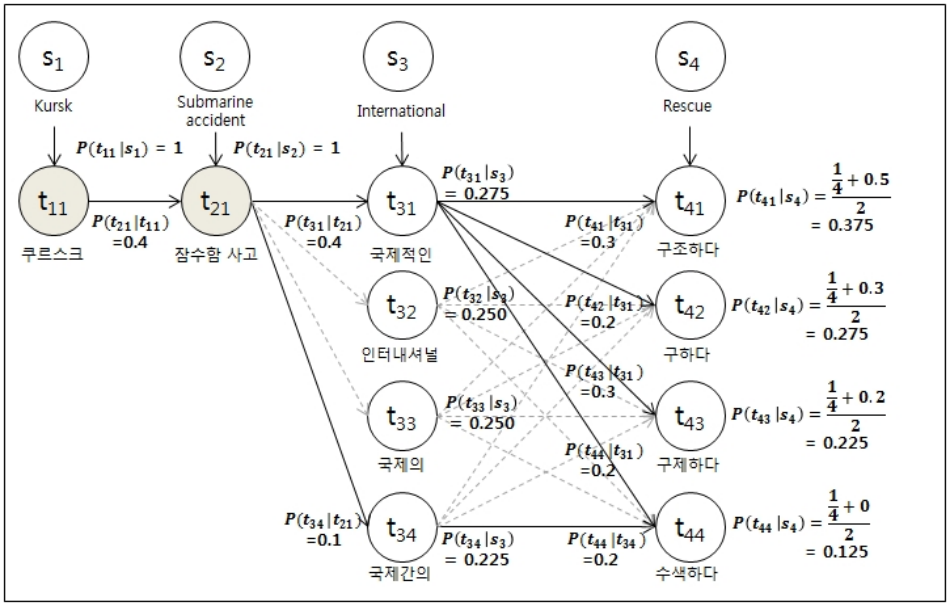


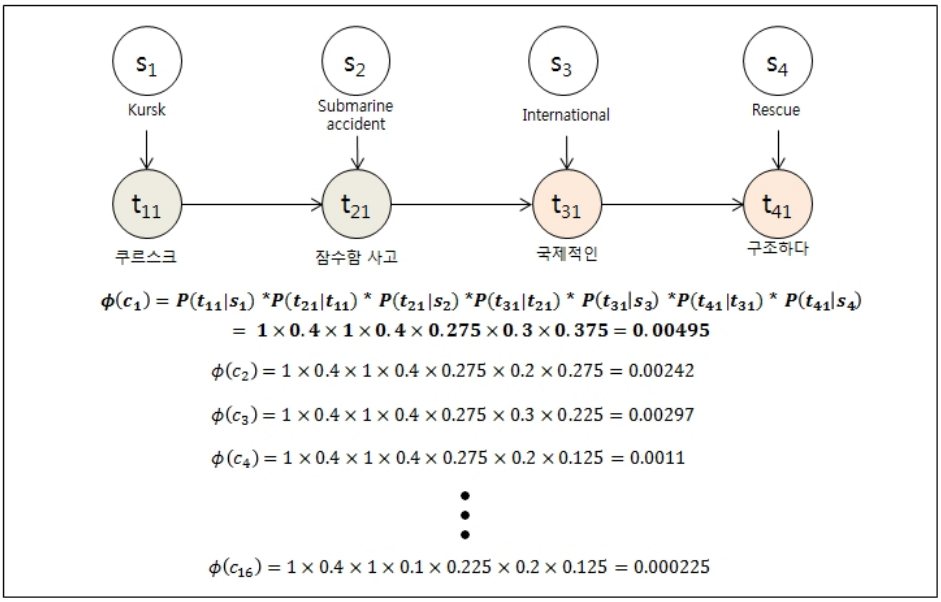
Figure 1. Overview of query translation

4.1. Translation Candidate Generation

We process an English query by first tokenizing, then removing English stopwords using the SMART stopword list, and then stemming. We then segment the translatable tokens or token sequences. We refer to the resulting tokens or token sequences and phrases as "terms." By "translatable," we mean that the term appears on the English side of the selected translation lexicons (WBL, MRD or CTL). Untranslatable terms are simply omitted, although translations of some untranslatable terms might be reinserted by post-translation query expansion (section 4.3). For example, the NTCIR-5 English query "Kursk, submarine accident, international rescue" is segmented using a longest-first greedy matching strategy "Kursk," "submarine accident," "international" and "rescue." For the CTL, this results in "Kursk => 쿠르스크," "submarine accident => 잠수함 사고," being translated using translations originally from the WBL, with "international => 국제적인, 인터내셔널, 국제의, 국제간의" and "rescue => 구조하다, 구하다, 구제하다, 수색하다" translated using translations originally from the MRD (see Figure 2). On average, 10% of query terms are untranslatable and thus omitted, but in this example every term happens to be translatable.



(a) full paths to generate 16 candidate target queries



(b) score calculation of candidate target-language queries and selection of best one

Figure 2. Example of the proposed query translation method using WBL and MRD. In (b), c_1 can be written by $c_1 = \{c_{11}(= t_{11}),\ c_{12}(= t_{21}),\ c_{13}(= t_{31}),\ c_{14}(= t_{41})\}$ as the notation of Equation 2.

4.2. Translation Selection

Some terms, most notably relatively common terms, can have many known translations, and effectiveness gains can often be obtained by giving some translations more weight than others. We propose a novel query translation method that generates all the candidate target-language (Korean) queries for the source-language (English) query Q_s, assigns a score to each, ranks the candidate target queries by that score, and then chooses the candidate target-language query with the top score.

$$Q_S = \{s_1, s_2, s_3, \ldots, s_n\} \tag{1}$$

$$\phi(c_i) = P(c_{i1} \mid s_1) \prod_{j=1}^{n-1} P(c_{ij+1} \mid c_{ij}) P(c_{ij+1} \mid s_{j+1}) \tag{2}$$

$$Q_t = \underset{c_i}{\operatorname{argmax}}\, \emptyset(c_i) \tag{3}$$

where, s_k is the k-th term of the source-language query Q_s, $\Phi(c_i)$ is the score of the i-th candidate target-language query c_i (In Figure 2, there are 16 candidate target-language queries from c_1 to c_{16}), c_{ij} is the j-th term of the candidate target query c_i (In Figure 2 (b), c_1 is written by $c_1 = \{c_{11}(= t_{11}),\ c_{12}(= t_{21}),\ c_{13}(= t_{31}),\ c_{14}(= t_{41})\}$), $P(c_{ij}|s_j)$ is the translation probability from s_j to c_{ij}, and $P(c_{ij+1}|c_{ij})$

is the transition probability (calculated by normalizing association scores using Equation 7). Candidate target-language queries are generated by combining possible translations of each query term. Figure 2 illustrates the calculation process of the scores ($\emptyset(c_i)$) for some candidate target-language queries. There are sixteen candidate target-language queries denoted by c_1={쿠르스크, 잠수함 사고, 국제적인, 구조하다}, c_2={쿠르스크, 잠수함 사고, 국제적인, 구하다}, c_3={쿠르스크, 잠수함 사고, 국제적인, 구제하다}, c_4={쿠르스크, 잠수함 사고, 국제적인, 수색하다}, … and c_{16}={쿠르스크, 잠수함 사고, 국제간의, 수색하다}, and their scores are calculated by Equation 2. The translation probabilities $P(c_{ij}|s_j)$ and $P(c_{ij+1}|c_{ij})$ are estimated by Equations 7 and 8 below, respectively. Eventually, c_1 is selected as the target-language query because it has the highest score, as shown in Figure 2.

To compute the association score α, we employ Jansen-Shannon Divergence, also known as Total Divergence to the Mean (TDM), which has been shown to be a useful measure for term association in other applications (Dagan et al., 1997). TDM measures the symmetrized Kullback-Leibler divergence to the mean of the two vectors.

$$\alpha(\overrightarrow{w_1}, \overrightarrow{w_2}) = 2\log 2 + \sum_{y \in both}\{\overrightarrow{w_1}(y) log \frac{\overrightarrow{w_1}(y)}{\overrightarrow{w_1}(y)+\overrightarrow{w_2}(y)} + \overrightarrow{w_2}(y) log \frac{\overrightarrow{w_2}(y)}{\overrightarrow{w_1}(y)+\overrightarrow{w_2}(y)}\} \quad (4)$$

where $y \in both$ means that both y-th element values of vectors $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$ are not 0. It should be noted that the sum of element values of each vector must be 1 in this equation. Vector $\overrightarrow{w_i}$ is defined as follows:

$$\overrightarrow{w_i} = (wt_{i1}, wt_{i2}, \dots, wt_{in}) \quad (5)$$

$$wt_{ij} = \frac{P(w_i|d_j)}{\sum_{k=1}^{n} P(w_i|d_k)}, \quad P(w_i|d_j) = \frac{tf_{ij}}{dl_j} \quad (6)$$

where vector $\overrightarrow{w_i}$ is the term distribution vector for the i-th term w_i, wt_{ij} is the weight of w_i in the j-th document, n is the number of documents in some collection that is used to calculate the term distribution, and $P(w_i|d_j)$ is the probability that the term w_i occurs in the j^{th} document d_j; tf_{ij} is the term frequency of w_i in d_j and dl_j is the length of document d_j. $\sum_{k=1}^{n} P(w_i|d_k)$ is the normalization factor

required to ensure that $\sum_{j=1}^{n} wt_{ij} = 1$.

Suppose that the *j+1*-st term s_{j+1} of source query Q_s has *m* translations $\{t_{j+1,1}, t_{j+1,2},\ldots, t_{j+1,m}\}$, and $c_{ij} = t_{j1}$ and $c_{ij+1} = t_{j+1,1}$; c_{ij} and c_{ij+1} are the *j-th and j+1*-st terms of candidate target-language query c_i respectively. In this case, the transition probabilities, $P(c_{ij+1}|c_{ij})$, are estimated by normalizing the association scores as follows:

$$\mathrm{P}\left(c_{ij+1} = t_{j+1,1} \middle| c_{ij} = t_{j1}\right) = \frac{\alpha(t_{j1}, t_{j+1,1})}{\sum_{k}^{m} \alpha(t_{j1}, t_{j+1,k})} \tag{7}$$

where $\alpha(t_{j1}, t_{j+1,1})$ is the association score between translations t_{j1} and $t_{j+1,1}$.

Translation probabilities $P(c_{ij}|s_j)$ are estimated from parallel text using translation probabilities from Giza++ tools, filtered, and optionally then smoothed, using the CTL. For Lexical Filtering (LF), the intersection between translations in the CTL and translations with non-zero (thresholded) translation probabilities is first computed, and finally the translation probabilities for those equivalents are renormalized by dividing by the sum of the remaining translation probabilities. We can in principle learn a translation relationship between any pair of words, almost always with very low probability. In order to achieve a useful degree of filtering, we therefore need to threshold the probabilities in some way. We therefore set a threshold on the Cumulative Distribution Function (CDF); with a threshold of 0 corresponding to using only the top-ranked query translation (a well-studied baseline) and a threshold of 1 corresponding to considering all alternatives. As we proceed down the list of candidate queries, as ranked by Equation 8, we add each new translation alternative with the probability for the translated query in which that translation alternative first appears. We stop when we reach the CDF threshold (in our experiments, 0.6) and we combine the results using the Indri's *#wsyn* operator (as in PSQ; see Section 5 below).

Because our parallel text is small, we have reason to suspect that the translation probabilities for relatively uncommon terms (precisely those terms we would expect to see in queries that are formed with specificity in mind) may be poorly estimated. Our Lexical Smoothing (LS) approach seeks to mitigate this by relaxing the estimated translation probabilities back toward a uniform distribution. We do this by first assigning all translations for each source-language CTL term a uniform distribution and then averaging those uniform probabilities with the normalized translation probabilities that had been computed using the LF method. For example, if there are *m* translations $\{t_{j1}, t_{j2},\ldots, t_{jm}\}$ of a

source language term s_j, the new translation probabilities of the target-language query term $c_{ij} = t_{jk}$ given the source-language query term s_i can be calculated as:

$$\mathrm{P}(c_{ij} = t_{jk}|s_j) = \frac{\left\{\frac{1}{m} + P_{giza}(t_{jk}|s_j)\right\}}{2} \quad (8)$$

where $P_{giza}(c_{ij} = t_{jk}|s_j)$ is translation probability from Giza++. Figure 2 shows how to use the translation probabilities computed using the LS method.

For notational convenience we refer to our technique for selecting among translations as WTDM (for Weighted Total Divergence from the Mean). When reporting experiment results, we denote the full process described here CLIR-LS-WTDM (for Cross-Language IR, Lexical Smoothing, association by Weighted Total Divergence to the Mean), CLIR-LF-WTDM or CLIR-CTL-WTDM.

4.3. Post-Translation Query Expansion

In this section we describe our method for post-translation query expansion using cross-article hyperlinks from Wikipedia. We first present an overview of the query expansion process (see Figure 3) and then describe each step in detail.

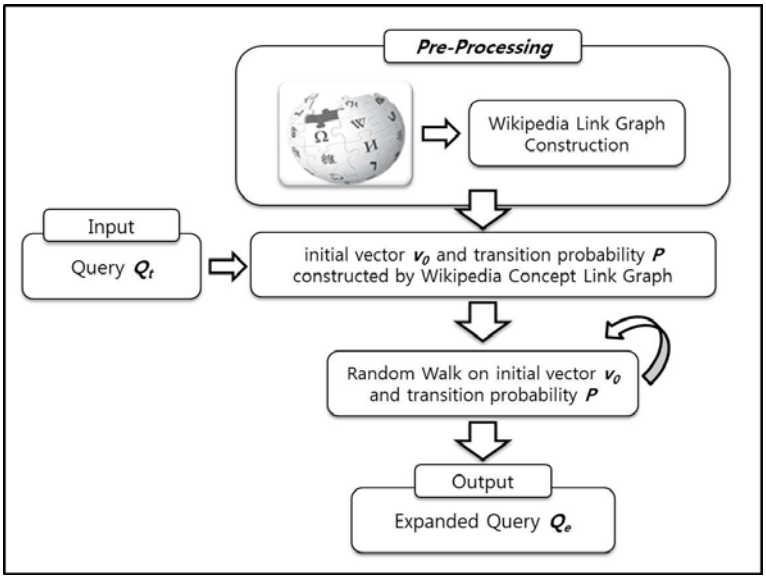
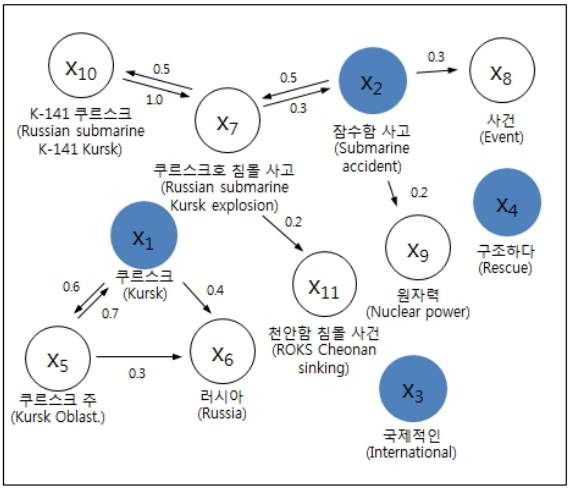


Figure 3. Overview of query expansion

Relationships between Wikipedia concepts are estimated using a random walk on the graph *G* of target-language Wikipedia pages described above (see Section 3.5). Basically, the walk starts at some

Wikipedia page corresponding to a term in the translated query and, at each step, moves to a neighboring Wikipedia page that is randomly chosen according to some distribution (Avin and Brito, 2004; Collins-Thompson and Callan, 1992; Hu et al., 2009).



(a) link graph G

P	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
x1	0	0	0	0	0.6	0.4	0	0	0	0	0
x2	0	0	0	0	0	0	0.5	0.3	0.2	0	0
x3	0	0	0	0	0	0	0	0	0	0	0
x4	0	0	0	0	0	0	0	0	0	0	0
x5	0.7	0	0	0	0	0.3	0	0	0	0	0
x6	0	0	0	0	0	0	0	0	0	0	0
x7	0	0.3	0	0	0	0	0	0	0	0.5	0.2
x8	0	0	0	0	0	0	0	0	0	0	0
x9	0	0	0	0	0	0	0	0	0	0	0
x10	0	0	0	0	0	0	1.0	0	0	0	0
x11	0	0	0	0	0	0	0	0	0	0	0

(b) transition probability matrix *P*

v_0		$P^T v_0$		v_1		$P^T v_1$		v_2		$P^T v_2$		v_3	
x1	1	x1	0	x1	0.50	x1	0.210	x1	0.3550	x1	0.2100	x1	0.2825
x2	1	x2	0	x2	0.50	x2	0.075	x2	0.2875	x2	0.0750	x2	0.1813
x3	1	x3	0	x3	0.50	x3	0	x3	0.2500	x3	0	x3	0.1250
x4	1	x4	0	x4	0.50	x4	0	x4	0.2500	x4	0	x4	0.1250
x5	0	x5	0.6	x5	0.30	x5	0.300	x5	0.3000	x5	0.2130	x5	0.2565
x6	0	x6	0.4	x6	0.20	x6	0.290	x6	0.2450	x6	0.2320	x6	0.2385
x7	0	x7	0.5	x7	0.25	x7	0.250	x7	0.2500	x7	0.2063	x7	0.2281
x8	0	x8	0.3	x8	0.15	x8	0.150	x8	0.1500	x8	0.0863	x8	0.1181
x9	0	x9	0.2	x9	0.10	x9	0.100	x9	0.1000	x9	0.0575	x9	0.0798
x10	0	x10	0	x10	0	x10	0.125	x10	0.0625	x10	0.1250	x10	0.0938
x11	0	x11	0	x11	0	x11	0.050	x11	0.0250	x11	0.0500	x11	0.0375

(c) weight vector v_j at iteration i_k

Figure 4. Post-translation query expansion example.

We first initialize the transition probability matrix *P* based on the link graph *G*. We define transition probabilities $P_{t+1|t}(x_k|x_j)$ from the vertex x_j to x_k ($x_j \in X$ *and* $x_k \in X$) by normalizing the score out of node x_j, so,

$$P_{t+1|t}(x_k|x_j) = \frac{score_{jk}}{\sum_i score_{ji}} \tag{9}$$

where i ranges over all vertices connecting to x_j and $score_{ij}$ denotes the score of link from x_i to x_j. The link score is estimated by the number of hyperlinks from x_i to x_j. The notation $P_{t+1|t}(x_k \mid x_j)$ denotes the transition probability from node x_j at time t to node x_k at time $t+1$. We rewrite the one-step transition probabilities in a matrix form as $P = [P_{t+1|t}(x_k \mid x_j)]_{jk}$ with a size of $m \times m$. The matrix P is row stochastic so the sum of the rows in P is 1.

We then initialize the term weight vector using the translated the query, as shown in the first vector in Figure 3(c). The initial vector v_0 is an m-dimension vector with values as follows:

$$v_0 = \{p_0(x_j)\}_{j=1}^{m} \text{ for } j = 1,2,\ldots,m \tag{10}$$

$$p_0(x_j) = \begin{cases} 1, if\, x_j \in Q \\ 0, otherwise \end{cases} \tag{11}$$

where $p_0(x_j)$, the probability that term i is in the query at the start of the random walk, is either 0 or 1.

Given this definition, the random walk algorithm works by repeatedly reweighting all terms, including potential query expansion terms, iteratively multiplying the transition probability matrix P by the current vector v_i. as follows:

Input : Transition probability matrix P and initial vector v_0
Output : v_n

1 : for i := 0 to n
2 : compute $v_{i+1} = \alpha P^T v_i + (1-\alpha)v_i$, where $\alpha \in [0,1)$
3 : return v_n
4: Choose top k terms from v_n for query expansion

where P^T is the transpose of P. For the experiments, we describe below (Section 7) we arbitrarily set α to 0.5, and we sweep k between 10 and 100 in steps of 10.

5. Baseline Techniques

We seek to show improvements over strong baselines, so we have implemented baselines for query translation and post-translation query expansion that are among the best known ways of using similar resources.

5.1. Query Translation Baselines

For comparison purposes, four baseline techniques were tried for query translation: Pirkola's Structured Queries (SQ), Seo's Total Divergence to the Mean (TDM) association measure, Darwish's Probabilistic Structured Queries (PSQ), and Google Machine Translation (MT).

SQ is the simplest of our four baseline query translation techniques, assuming only that some (unweighted set) of possible translations is known for each query term. The key idea is to estimate term frequency and document frequency for each query term using document-language evidence, and then to compute term weights in the query language (Prikola, 1998). This is well known to be a strong baseline when translation probabilities are not known.

When translation probabilities are not available, the most widely used alternative to SQ is one-best query translation using some form of association measure among the translated query terms. We implement this baseline by using the Total Divergence to the Mean (TDM) term association measure (i.e., Jensen-Shannon Divergence) in the manner described by Seo et al. (2005). Seo et al. calculated TDM association scores among every term pair in the Cartesian product of possible query translations using Equation 4 above (see Section 4.2) and then made the one-best selection for the entire translated query that maximizes (over all translated queries) the sum of the pairwise term association scores for every term pair in the translated query. TDM differs from our WTDM approach in two ways. First, TDM uses the sum of the association measures, whereas WTDM uses the product of the association measures. The result of this difference is that TDM would prefer a more sharply skewed distribution of association scores if it results in a higher sum, whereas WTDM would prefer a better balanced distribution of association scores if it results in a higher product. In other words, WTDM is biased more strongly against even a single very low association score. Second, TDM takes no account of translation probabilities, thus rewarding strongly associated translations even if there is reason to believe that one or both of those translations would be very unlikely to be correct.

PSQ is a straightforward extension to SQ in which translation probabilities are used to weight the term frequency and document frequency estimates. The key idea in PSQ is that translation

probabilities should be applied to scale term frequency and document frequency statistics before computing term weights (i.e., "translating" counts rather than weights). When reasonably accurate translation probabilities are available, PSQ has been observed to yield better CLIR ranked retrieval effectiveness than SQ. We threshold the Cumulative Distribution Function (CDF) at 0.6 to limit the effect of spurious term translation alignments, which has been found to be a reasonable approach in prior work (Wang and Oard, 2006).

Statistical machine translation is, in essence, just another way of using translation probability statistics. We use the Google MT system for query translation, which has the advantage of leveraging the very large collection of Korean-English parallel text that is available to Google. A potential disadvantage of MT, however, is its reliance on a document-tuned (rather than a query-tuned) language model.

5.2. Post-Translation Query Expansion Baselines

Most reported work on query expansion draws on the idea of Blind Relevance Feedback (BRF) in which the key idea is to find discriminative terms in highly ranked queries and add them to the query in an effort to include a broader range of related terms. This can often improve retrieval effectiveness (on average, over many topics) when used with large collections in which the initial query has a good chance of finding a few highly-ranked relevant documents. We implement BRF using Robertson and Walker's relevance weights (Robertson and Walker, 1999). In this technique, every term that occurs in more than one document among top *R* documents (in our case, our Korean NTCIR collection, as retrieved by Indri) is sorted according to its relevance weight, which is calculated as:

$$RW(t) = r_t \log\frac{N}{n_t} - \log\binom{R}{r_t} - \log V \tag{12}$$

where *RW(t)* is the relevance weight of term *t*, *R* is the total number of top-ranked documents, r_t is the number of top-ranked documents in which the term *t* occurs, *N* is the size of the collection, n_t the number of documents that contain the term *t*, and *V* is the size of the vocabulary. The argument of the second logarithm is the number of ways one can choose r_t from *R*. We set *R* to 10, which is a

commonly used value. In addition, we implemented the other version of BRF using the Korean Wikipedia collection instead of the Korean NTCIR collection. It is one of comparison techniques to utilize the Wikipedia collection for query expansion. For clarity, we refer to BRF using the same collection that we are searching as Conventional BRF (CBRF). We also implemented the same method for BRF, but replacing the Korean NTCIR collection with Korean Wikipedia as the expansion collection; we refer to this as Wikipedia BRF (WBRF).

Because the link structure of Wikipedia offers an alternative basis for query expansion that we exploit in our own work, we also implemented the technique of Elsas et al (2008) for leveraging cross-article hyperlinks in Wikipedia to acquire richer vocabulary; this is named by WLS (Wikipedia Link Structure) in our experiments. The key idea is to search Wikipedia for related documents, and then to take as expansion terms not terms from the title of highly ranked Wikipedia articles found during the expansion process, but rather terms from the anchor text in other related Wikipedia articles that link to those articles This link-based query expansion technique begins by using the original query to search Wikipedia. From the resulting ranking of Wikipedia articles, two sets are defined; the relevant set, *SR*, is the articles ranking in the top *R*, and the working set, *SW*, is the articles ranking in the top *W* (for *W*>*R*). Following prior work, we set *R*=100 and *W*=1,000 (Elsas et al., 2008). Each anchor phrase a_i occurring in an article in *SW* and linking to an article in *SR* is then scored as follows:

$$\text{score}(a_i) = \sum_{a_{ij} \in SW \text{ and } anchored\_article(a_{ij}) \in SR} (R - rank(anchored\_article(a_{ij}))) \quad (13)$$

where a_{ij} denotes an occurrence of anchor phrase a_i in the working set and $anchored\_article(a_{ij})$ is the article in the relevant set that is linked to by the hyperlink of a_{ij}. We then simply choose the top *k* ranked terms as before (sweeping *k* from 10 to 100 in steps of 10) and add them to the query.

6. Experiment Design

We tested our proposed query translation and query expansion techniques using the NTCIR-5 English-Korean CLIR test collection. In this section, we describe the design of our experiments.

6.1. Test Collection

The NTCIR-5 English-Korean CLIR test collection that we used contains 220,374 Korean documents

and 50 topics.[8] The document set consists of Hankookilbo and Chosunilbo newspaper articles published in 2000 and 2001, and each topic consists of title, description, narrative, and concept fields (Kwok et al., 2005). In our experiments, two types of queries were used: title field only (*Title*), and title and description fields (*Title+Description*). In this paper, we focus on results for *Title* queries because the results are similar for *Title+Description* queries. For reference, the *Title+Description* results are shown in the Appendix.

There are two interpretations of relevance that are commonly reported for the NTCIR-5 test collection: rigid (in which only highly relevant documents are scored as relevant) or relaxed (in which both highly relevant documents and somewhat relevant documents are scored as relevant). Relaxed relevance corresponds to the standard for relevance judgment in the widely cited Text Retrieval Conferences (TREC), and we report only relaxed relevance. We did, however, also run experiments with rigid relevance, obtaining broadly similar results.

6.2. Evaluation Measures

We report three evaluation measures: Mean Average Precision (MAP), mean 11-point average precision (AP11), and mean R-precision (R-Precision). MAP is a widely reported measure that characterizes the degree to which a ranked list places relevant documents ahead of others, with particular emphasis on the ranking near the top of the list (so-called "early precision"). AP11 is an interpolated predecessor to MAP that was commonly reported when test collections were smaller; it is reported here for comparison with Seo et al. (2005). R-Precision is often reported as an alternative measure when seeking to characterize retrieval results in a way that is easily explained; it is computed as the fraction of the retrieved documents that are relevant at the first point in the ranked list where every relevant document could have been retrieved. All three measures are known to be fairly highly correlated (in the way they rank systems), and all three can reasonably be averaged over topics that have markedly different characteristics (in contrast, for example, to precision at 10, which does not average meaningfully when some topics have very few relevant documents and others have very many). We report statistical significance for observed differences in MAP when the p value is sufficiently small ($p<0.01$ or $p<0.05$) by a one-sided paired t-test.[9]

[8] http://research.nii.ac.jp/ntcir/index-en.html
[9] http://www.graphpad.com/quickcalcs/ttest1.cfm

6.3. Information Retrieval System

We indexed both the Korean NTCIR-5 articles and the Korean Wikipedia articles using Indri, an information retrieval system from the University of Massachusetts. Indri supports an extensive set of query operators based on the earlier Inquery system's query language (Callan et al., 1992) that can support both SQ and PSQ when using a language model for ranking. For all of the methods that we implemented, the construction of the final query, Q_{final}, is illustrated by the following Indri query template.

$$Q_{final} = \#\text{weight}(\lambda_{fb}\ Q_{base}\ \ (1-\lambda_{fb})\ Q_{exp}) \tag{14}$$

The initial query, Q_{base} for every technique other than PSQ is formulated as follows:

$$Q_{base} = \#\text{combine}\left(\#\text{syn}(t_{11},\cdots t_{1j})\ \#\text{syn}(\text{t}_{21},\cdots \text{t}_{2j})\cdots\ \#\text{syn}(\text{t}_{i1}\ t_{i2},\cdots \text{t}_{ij})\right) \tag{15}$$

where t_i are source-language query terms or phrases, and t_{ij} are the translations of t_i. For PSQ, we replace Indri's *#syn* operator with Indri's *#wsyn* operator, and we provide the translation probability as the weight for each translation in the *#wsyn* operator.

Q_{exp} is a weighted query of the form

$$Q_{exp} = \#\text{weight}(\gamma_1\#\text{combine}(e_1)\ \gamma_2\#\text{combine}(e_2)\ \cdots\gamma_T\#\text{combine}(e_T)) \tag{16}$$

where e_i are expansion terms or phases and γ_i are the weights assigned by the query expansion algorithm.

In all of our experiments that use feedback, the feedback mixing weight, λ_{fb} is fixed at 0.5, the number of feedback documents (where applicable) was set to 10, and the number of expansion terms for each technique was selected as the value yielding the best MAP results on a grid search from 10 to 100 with step size 10 (this value was 30 expansion terms for Mono-CBRF, CLIR-MRD-TDM-CBRF, CLIR-CTL-TDM-RW and CLIR-CTL-LS-WTDM-RW, 50 expansion terms for CLIR-CTL-TDM-

WBRF, and 20 expansion terms for CLIR-CTL-TDM-WLS).

7. Results

In this section, we present results from several experiments that demonstrate the effectiveness of our proposed techniques.

7.1. Results for Lexical Filtering (LF) and Lexical Smoothing (LS)

Table 7 shows the results for Lexical Filtering (LF) and Lexical Smoothing (LS), along with comparable results for five baselines. The strongest baseline, Mono-CBRF, results from using same-language (Korean) queries and performing blind relevance feedback. For ease of comparison, results for other conditions are reported both as absolute measures (for AP11, MAP, and R-Precision) and as fractions of the value obtained by Mono-CBRF for each measure that was achieved by each technique. As comparison with the Mono condition shows, CBRF contributes about a 10% improvement in the monolingual condition, which is consistent with previously reported results. The strongest cross-language baseline (using English queries) is the CLIR-MT condition in which the queries were translated from English to Korean using Google. As the CLIR-PSQ condition shows, the using our small parallel text collection, far smaller than that available to Google, yields quite poor results – worse even than our CLIR-CTL-SQ baseline that makes no use of translation probabilities. Figure 5 illustrates this comparison graphically. We see a similar effect when we compare CLIR-CTL-LF-PSQ (which uses the same probabilities as CLIR-PSQ) with CLIR-CTL-LS-PSQ (in which those translation probabilities are relaxed towards a uniform distribution): LS is markedly better than LF. Comparing CLIR-CTL-LS-PSQ with CLIR-CTL-SQ indicates, however, that translation probabilities from a small parallel text collection are indeed useful, since SQ (which has no access to translation probabilities) does markedly worse than LS. From this we conclude that our lexical smoothing technique provides a useful way of combining lexical and statistical evidence, and that it can be useful to make such a combination when only a limited amount of parallel text is available.

Table 7. Experiments with Translation Probability Smoothing

Technique	Query Type	AP11	MAP	R-Prec
Mono-CBRF	Title	0.3278	0.4029	0.3548
Mono	Title (% Mono-CBRF)	0.2986 (91%)	0.3636 (90%)	0.3158 (89%)
CLIR-MT	Title (% Mono-CBRF)	0.2296 (70%)	0.2870 (71%)	0.2547 (72%)
CLIR-CTL-LS-PSQ	Title (% Mono-CBRF)	**0.1834** **(56%)**	**0.2314** **(57%)**	**0.2265** **(64%)**
CLIR-CTL-SQ	Title (% Mono-CBRF)	0.1241 (38%)	0.1530 (38%)	0.1384 (39%)
CLIR-CTL-LF-PSQ	Title (% Mono-CBRF)	0.0897 (27%)	0.1174 (29%)	0.0984 (28%)
CLIR-PSQ	Title (% Mono-CBRF)	0.0656 (20%)	0.0894 (22%)	0.0786 (22%)

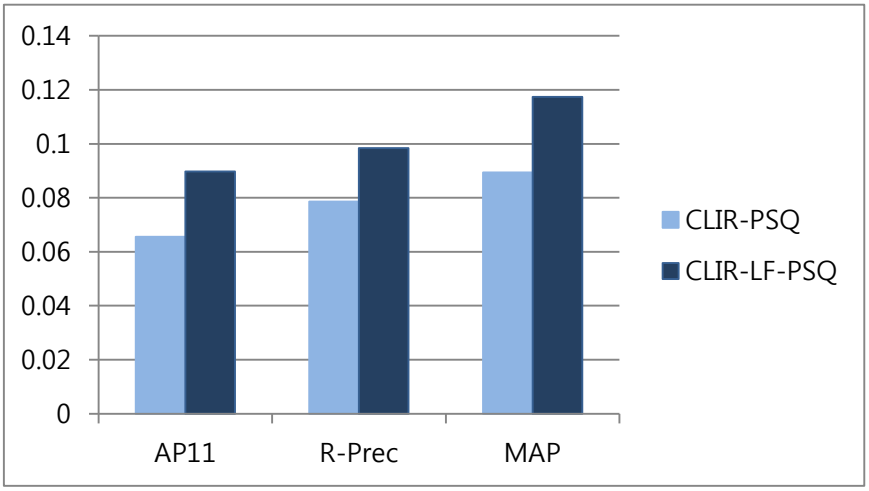


Figure 5. Comparing the Effectiveness of CLIR-PSQ *vs.* CLIR-LF

7.2. Results for Weighted Total Divergence to the Mean (WTDM)

Table 8 compares three ways of using association evidence. As expected, the large CTL, which includes both the Machine-Readable Dictionary (MRD) and the Wikipedia-Based Lexicon (WBL) yields better results than can be obtained with the smaller MRD alone when used with the TDM baseline technique for association evidence, and this difference is statistically significant ($p<0.01$).[10] Figure 6 illustrates this comparison. TDM cannot take advantage of translation probabilities, however, but WTDM can. As the comparison between CLIR-CTL-LS-WTDM and CLIR-CTL-TDM shows, translation probability evidence helps considerably, yielding an 8% relative improvement in MAP that is statistically significant ($p<0.05$). Indeed, comparing Tables 7 and 8 for comparable conditions, we see that TDM is markedly better than SQ and that WTDM is markedly better than PSQ. Moreover,

[10] We test statistical significance on differences in Mean Average Precision (MAP) using a one-sided *t*-test.

CLIR-CTL-LS-WTDM even does somewhat better than CLIR-MT despite having access to a far smaller collection of parallel text.

Table 8. Experiments with TDM and WTDM Association Evidence

Technique	Query Type	AP11	MAP	R-Prec
CLIR-CTL-LS-WTDM	Title (% Mono-CBRF)	**0.2708** **(83%)**	**0.3619** **(90%)**	**0.3051** **(86%)**
CLIR-CTL-TDM	Title (% Mono-CBRF)	0.2499 (76%)	0.3364 (83%)	0.2802 (79%)
CLIR-MRD-TDM	Title (% Mono-CBRF)	0.1858 (57%)	0.2401 (60%)	0.2113 (60%)

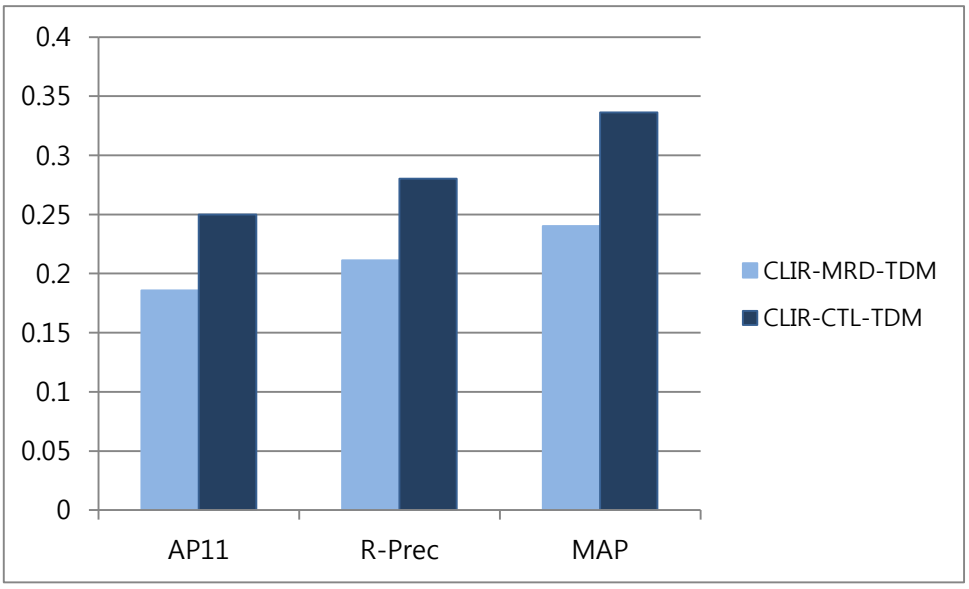


Figure 6. Comparing the Effectiveness of CLIR-MRD-TDM *vs*. CLIR-CTL-TDM

7.3. Results for Post-Translation Query Expansion

Table 9 compares two baseline query expansion techniques with our proposed Random Walk (RW) method. As can be seen, CLIR-CTL-TDM-WLS (which uses the Wikipedia link structure) yields about the same results as CLIR-CTL-TDM-WBRF. Comparing CLIR-CTL-TDM-RW with CLIR-CTL-TDM-WLS shows that our RW technique is considerably more effective. This difference is statistically significant ($p<0.01$). Comparing CLIR-CTL-LS-WTDM-RW with CLIR-CTL-TDM-RW shows that WTDM yields a further 3% improvement that is statistically significant ($p<0.05$). This is a smaller improvement than we observed without post-translation expansion, thus confirming that (as expected) post-translation query expansion tends to mitigate some weaknesses in the query translation process. Comparing Tables 8 and 9, we see a very substantial benefit from using RW for post-

translation query expansion, with large and statistically significant MAP differences between CLIR-CTL-TDM-RW and CLIR-CTL-TDM and between CLIR-CTL-LS-WTDM-RW and CLIR-CTL-LS-WTDM (both at $p<0.01$). Indeed, both CLIR-CTL-LS-WTDM-RW and CLIR-CTL-TDM-RW outperform even Mono-CBRF, which is consistent with the large and statistically significant difference observed between CLIR-CTL-TDM-RW and CLIR-CTL-TDM-WBRF. From this, we conclude that using a random walk on the Wikipedia link graph is an excellent approach to query expansion, and that application of the same technique in monolingual applications merits investigation in future work.

Table 9. Effectiveness of the Query Expansion methods

Technique	Query Type	AP11	MAP	R-Precision
CLIR-CTL-LS-WTDM-RW	Title (%Mono-CBRF)	**0.3816 (116%)**	**0.4834 (120%)**	**0.4193 (118%)**
CLIR-CTL-TDM-RW	Title (% Mono-CBRF)	0.3694 (113%)	0.4675 (116%)	0.4008 (113%)
CLIR-CTL-TDM-WLS	Title (% Mono-CBRF)	0.3008 (92%)	0.3573 (89%)	0.3079 (87%)
CLIR-CTL-TDM-WBRF	Title (% Mono-CBRF)	0.2926 (89%)	0.3517 (87%)	0.3066 (86%)
CLIR-MRD-TDM-CBRF	Title (% Mono-CBRF)	0.2517 (77%)	0.3381 (84%)	0.2874 (81%)

Traditional BRF techniques such as CBRF, and even WBRF, can improve results for queries that find a few relevant highly-ranked documents with the initial query, but some queries (e.g., those that find only mid-ranked relevant documents) can be hurt, and others (e.g., those where most of the highly ranked documents that are initially retrieved are relevant) are neither helped much nor hurt much. Our RW technique adversely affects the results for only 10 of the 50 topics in the NTCIR-5 test collection and the adverse effect on Average Precision (AP) for those 10 topics is very small (on average, -0.0094 absolute). By contrast, our RW technique yields far larger improvements in AP (on average, +0.1582 absolute) for 39 topics. Only one topic that yields the same AP with or without RW has no improvement by RW since its Korean query translated by CLIR-CTL-LS-WTDM does not have any matched terms with Wikipedia titles. It happens in the case that source English query terms are translated by only MRD. RW eventually cannot be applied to expand this query. Figure 7 shows

the five largest uninterpolated Average Precision (AP) improvements from CLIR-CTL-LS-WTDM to CLIR-CTL-LS-WTDM-RW, and Figure 8 shows the five largest AP performance reductions; Table 10 shows the corresponding terms from the original title query. These two sets of queries also exhibit markedly different coverage statistics in our Wikipedia-Based Lexicon (WBL): about 90% of the terms in the initial queries for the five topics with the greatest improvement from RW are present on the source-language (English) side of the WBL, as compared with less than 50% of the terms in initial queries for the five topics that showed the least improvement. This makes sense, as queries with better coverage in the Wikipedia provide a better starting point for the random walk.

Table 10. Initial queries with largest AP improvements and performance reductions from Random Walk expansion

5 Largest Improvements in AP		5 Largest performance reductions in AP	
Topic 14	nanotechnology	Topic 07	Wen Ho Lee Case, classified information national security
Topic 22	mad cow disease	Topic 26	donation, millionaire, heritage
Topic 24	economy class, syndrome, flight	Topic 27	longevity, secret, Antonio Toddy
Topic 36	remote operation, robot	Topic 28	Bubka, human bird, retirement
Topic 47	Korean general election, 2000, HanNara Party	Topic 48	genetically engineered food, regulation

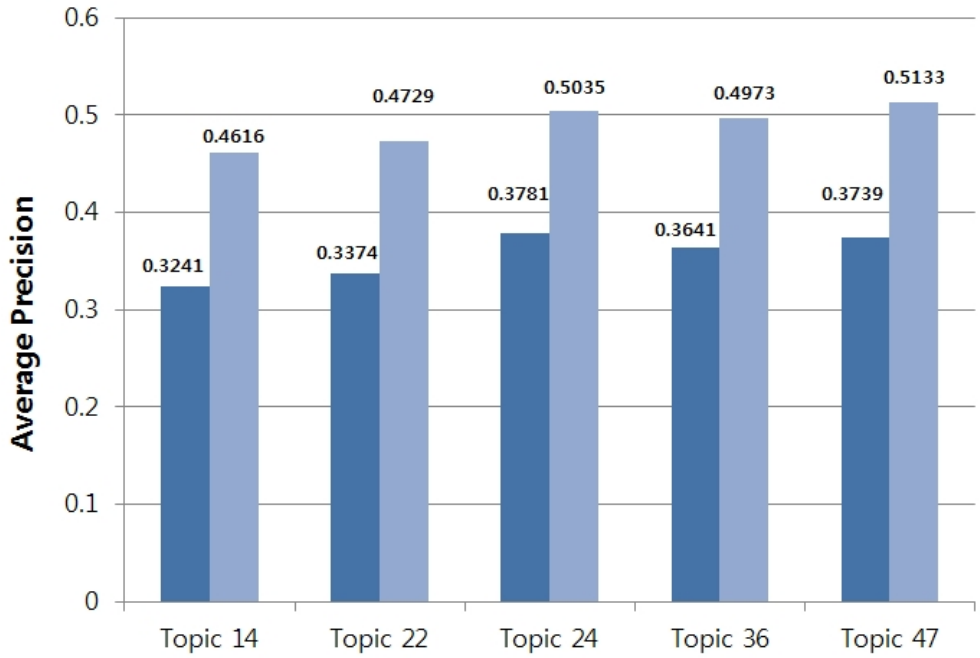


Figure 7. Largest AP differences of performance reductions between CLIR-CTL-LS-WTDM and CLIR-CTL-LS-WTDM-RW.

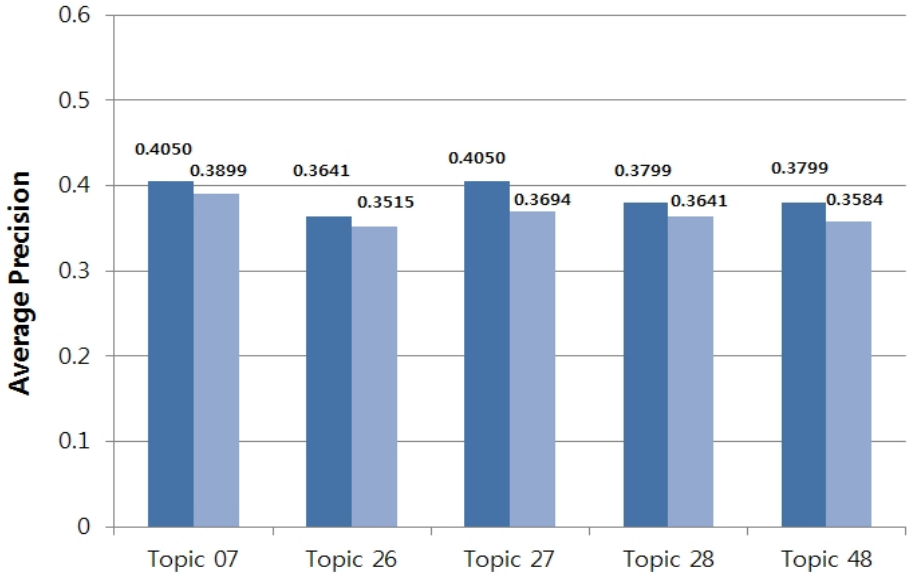


Figure 8. Largest AP differences of performance reductions between CLIR-CTL-LS-WTDM and CLIR-CTL-LS-WTDM-RW.

Figure 9 shows a topic-by-topic comparison of the effect of our RW techniques for post-translation query expansion. As can be seen, many topics exhibit large gains, and those few topics that are adversely affected (as measured by average precision) are not hurt much. Inspection of specific queries suggests that where an adverse effect is present, it is often because some query terms are not found in Wikipedia. Indeed, the one topic where the lines cross (i.e., where RW makes no difference at all) is the single case in which no query term was found in Wikipedia.

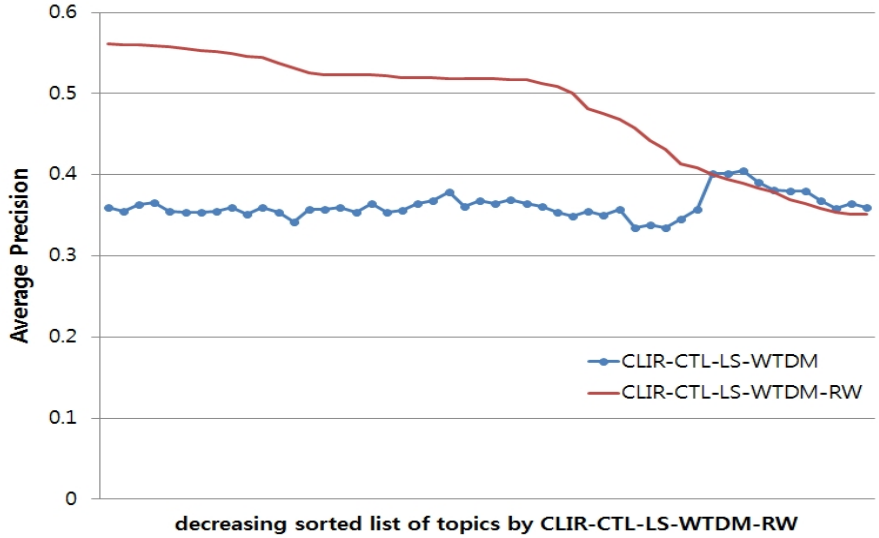


Figure 9. Topic-by-topic comparison showing AP with or without RW post-translation query expansion. 50 Topics are shown, plotted in order of decreasing AP for CLIR-CTL-LS-WTDM-RW.

7.4. Result Summary

Figure 10 summarizes our results. Early work on CLIR employed Machine Readable Dictionaries and association measures that lacked access to translation probabilities. Our CLIR-MRD-TDM baseline is representative of the results obtained by such methods. Adding translation knowledge (to create our CTL) and from parallel text (using our LS method to relax the learned translation probabilities toward a uniform distribution) yields large improvements in MAP that are statistically significant at $p<0.01$. Using the Korean Wikipedia link graph as a basis for post-translation query expansion yields further substantial improvements that are again statistically significant at $p<0.01$. This final result, with MAP above 0.48, achieves 120% of a strong monolingual baseline (Mono-CBRF), and this is currently the best reported result for English-Korean CLIR on the NTCIR-5 test collection.

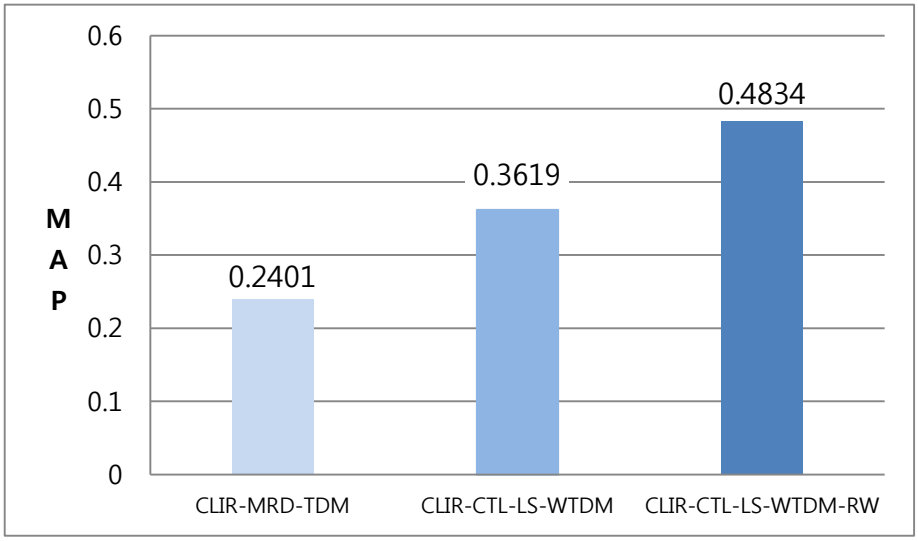


Figure 10. Net improvement over the CLIR-MRD-TDM baseline.

8. Conclusions and Future Work

The results we have reported make three contributions to CLIR research. First, we have shown that using a machine readable dictionary, a parallel corpus, and translation evidence from Wikipedia together can yield better results from any one source alone, or from any combination of two such sources. Second, we have shown how prior work on the application of Jensen-Shannon divergence to perform translation selection in a query translation architecture can be extended to instead perform translation weighting in ways that can yield improved ranked retrieval effectiveness. Third, we have

shown that the Wikipedia concept link graph can be used as a basis for effective post-translation query expansion.

It has long been known that large parallel text collections offer a useful source of evidence for CLIR. Until now, however, it has not been clear how smaller parallel test collections might be used productively. We have shown that by using lexical smoothing together with a term association technique that can leverage translation probabilities we can benefit from a small parallel text collection of a size that might be constructed even for low-resource languages. It has long been known that post-translation query expansion can be useful. Until now, however, improvements obtained on some queries came at the cost of (typically smaller) adverse effects on other queries. We have shown that using a random walk on the Wikipedia link graph can yield more consistent improvements, even for a language like Korean where the number of Wikipedia pages is far smaller than it is for English. And finally, we have shown that these improvements are present not only when using a machine-readable dictionary, but also when the coverage is augmented using translation, synonymy and polysemy relations extracted from Wikipedia.

As with any research, answering these questions helps us to ask new ones. McNamee and Mayfield (2002) have shown that the relative importance of different techniques can vary substantially with the size and richness of the available linguistic resources, so one productive direction for future work would be to repeat this study for additional language pairs for which the available translation and expansion resources are either larger or smaller. Another potentially interesting direction for future work would be to explore alternative design choices. For example, at present we assign each term the probability associated with the most likely translation of the entire query in which it appears. Alternative designs in which translation probabilities are accumulated over every full query translation in which a specific translated term appears would also be worth exploring. Similarly, in our present random walk algorithm, we either add an expansion term to a translated query or we do not. Weighted variants in which the weight decays as we follow additional links might also be worth exploring. These and other questions suggest that much still remains to be done to fully investigate how new resources such as Wikipedia can be productively used together with what came before to obtain better results that could be obtained with any one resource type alone.

Acknowledgement

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-013-D00116).

References:

Avin, C. and Brito, C. (2004). Efficient and robust query processing in dynamic environments using random walk techniques. In *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks (IPSN)*, pp. 277–286.

Ballesteros, L. and Croft, W.B (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Callan, J.P., Croft, W.B. and Harding, S.M. (1992). The INQUERY information retrieval system. In *Proceedings of the International Conference on Database and Expert Systems Applications*, pp. 78-83.

Collins-Thompson, K. and Callan, J. (2005). Query expansion using random walk models. In *Proceedings of 14th International Conference on Information and Knowledge Management*.

Dagan, I., Lee, L. and Pereira, F. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics*.

Darwish, K. and Oard, D.W. (2003). Probabilistic structured query methods. *In 26th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*.

Demner-Fushman, D. and Oard, D.W. (2003). The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *Hawaii International Conference on System Sciences*.

Elsas, J.L., Arguello, J., Callan, J. and Carbonell, J.G. (2008). Retrieval and feedback models for blog feed search. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 347-354.

Fautsch, C. and Savoy J. (2008). UniNE at TREC 2008: Fact and opinion retrieval in the blogsphere. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*.

Federico, M. and Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Gaillard, B., Boualem, M. and Collin, O. (2010). Query translation using Wikipedia-based resources for analysis and disambiguation. In *Proceedings of European Association for Machine Translation*.

Hu, J., Wang, G., Lochovsky, F., tao Sun, J. and Chen Z. (2009). Understanding user's query intent with Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, pp. 471-480.

Jones, G., Sakai, T., Collier, N., Kumano, A. and Sumita, K. (1999). A comparison of query translation methods for English–Japanese cross-language information retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 269-270.

Kwok, K.L., Choi, S., Dinstl, N. and Deng, P. (2005) NTCIR-5 Chinese, English, Korean cross language information retrieval experiments using PIRCS. In *Proceedings of NTCIR-5 Workshop Meeting, pp. 6-9*.

Laudauer, T.K. and Littman, M .L. (1990). Fully automatic document retrieval using latent semantic indexing. In *Proceedings of the annual conference of the UW Centre for the Oxford English Dictionary*.

Levow,, G.-A., Oard, D.W. and Resnik, P. (2005). Dictionary-Based Cross-Language Retrieval, *Information Processing and Management*, 41(3)523-547.

McCarley, J.S. (1999). Should we translate the documents or the queries in cross-language information retrieval. In *27th Annual Meeting of the Association for Computational Linguistics*.

McNamee, P. (2008). *Textual Representations for Corpus-Based Bilingual Retrieval*, Ph.D., Dissertation, Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, Baltimore, MD.

McNamee, P. and Mayfield, J. (2002). Comparing cross-language query expansion techniques by Degrading Translation Resources. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*, pp. 11-15.

Mishne, G. (2007). *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, Amsterdam.

Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, R., Hiemstra, D. and de Jong, F. (2008). WikiTranslate: Query translation for cross-lingual information retrieval using only Wikipedia. In *Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum*.

Oard, D.W. and Diekema, A. (1998). Cross-language information retrieval. In M. Williams (ed.), *Annual Review of Information Science*, pp. 223-256.

Oard, D.W. (2009). Multilingual information access, In *Encyclopedia of Library and Information Science* (3rd edition), Taylor and Francis.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-Language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pu, W., Jian,, H. Hua-Jun, Z. and Zheng, C. (2007). Improving text classification by using encyclopedia knowledge. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM-07)*.

Robertson, S.E., and Walker, S. (1999). Okapi/Keenbow at TREC-8. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, pp. 151-161.

Schönhofen, P., Benczúr, A., Bíró, I. and Csalogány, K. (2008). Cross-language retrieval with Wikipedia. *Advances in Multilingual and Multimodal Information Retrieval*, pp. 72-79.

Seo, H., Kim, S., Rim, H. and Myaeng, S. (2005). Improving query translation in English-Korean cross-language information retrieval. *Information Processing & Management*, Vol. 41, pp. 507-522.

Sorg and Cimiano (2008). Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop*.

Tang, L.-X., Trotman, A., Geva, S. and Xu, Y. (2010). Wikipedia and Web document based query translation and expansion for cross-language IR. In *Proceedings of NTCIR-8 Workshop Meeting*, pp. 121-125.

Wang, J. and Oard, D.W. (2006). Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*, pp. 202-209.

Zhang, W. and Yu, C. (2006). UIC at TREC 2006 blog track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*.

Appendix

This appendix presents the performances of the *Title+Description* queries.

Table A.1 Experiments with Translation Probability Smoothing

Technique	Query Type	AP11	MAP	R-Prec
Mono-CBRF	Title+Desc	0.3411	0.4227	0.3736
Mono	Title+Desc (% Mono-CBRF)	0.3117 (91%)	0.3841 (91%)	0.3395 (91%)
CLIR-MT	Title+Desc (% Mono-CBRF)	0.2436 (71%)	0.3089 (73%)	0.2611 (70%)
CLIR-CTL-LS-PSQ	Title+Desc (% Mono-CBRF)	**0.1927** **(56%)**	**0.2438** **(58%)**	**0.2419** **(65%)**
CLIR-CTL-SQ	Title+Desc (% Mono-CBRF)	0.1385 (41%)	0.1714 (41%)	0.1526 (41%)
CLIR-CTL-LF-PSQ	Title+Desc (% Mono-CBRF)	0.0948 (28%)	0.1264 (30%)	0.1114 (30%)
CLIR-PSQ	Title+Desc (% Mono-CBRF)	0.0731 (21%)	0.1044 (25%)	0.0934 (25%)

Table A.2 Experiments with TDM and WTDM Association Evidence

Technique	Query Type	AP11	MAP	R-Prec
CLIR-CTL-LS-WTDM	Title+Desc (% Mono-CBRF)	**0.2891** **(85%)**	**0.3821** **(90%)**	**0.3089** **(83%)**
CLIR-CTL-TDM	Title+Desc (% Mono-CBRF)	0.2741 (80%)	0.3642 (86%)	0.2923 (78%)
CLIR-MRD-TDM	Title+Desc (% Mono-CBRF)	0.1887 (55%)	0.2574 (61%)	0.2074 (56%)

Table A.3 Effectiveness of the Query Expansion methods

Technique	Query Type	AP11	MAP	R-Precision
CLIR-CTL-LS-WTDM-RW	Title+Desc (%Mono-CBRF)	**0.3866** **(113%)**	**0.4995** **(118%)**	**0.4229** **(113%)**
CLIR-CTL-TDM-RW	Title+Desc (% Mono-CBRF)	0.3756 (110%)	0.4883 (116%)	0.4084 (109%)
CLIR-CTL-TDM-WLS	Title+Desc (% Mono-CBRF)	0.3168 (93%)	0.3625 (86%)	0.3193 (85%)
CLIR-CTL-TDM-WBRF	Title+Desc (% Mono-CBRF)	0.3075 (90%)	0.3615 (86%)	0.3212 (86%)
CLIR-MRD-TDM-CBRF	Title+Desc (% Mono-CBRF)	0.2589 (76%)	0.349 (83%)	0.2988 (80%)