Dictionary-Based Techniques for Cross-Language Information Retrieval

Gina-Anne Levow,* Douglas W. Oard,† and Philip Resnik‡§

Abstract

Cross-language information retrieval (CLIR) systems allow users to find documents written in different languages from that of their query. Simple knowledge structures such as bilingual term lists have proven to be a remarkably useful basis for bridging that language gap. A broad array of dictionary-based techniques have demonstrated utility, but comparison across techniques has been difficult because evaluation results often span only a limited range of conditions. This article identifies the key issues in dictionary-based CLIR, develops unified frameworks for term selection and term translation that help to explain the relationships among existing techniques, and illustrates the effect of those techniques using four contrasting languages for systematic experiments with a uniform query translation architecture. Key results include identification of a previously unseen dependence of pre-translation and post-translation expansion on orthographic cognates and development of a query-specific measure for translation fanout that helps to explain the utility of structured query methods.

Keywords: Cross-language Information Retrieval, Ranked retrieval, Dictionary-based translation.

1 Introduction

In the book of Genesis, the following passage describing the impact of linguistic diversity on mankind's ability to create great works (in this case, the Tower of Babel) seems particularly apt to the situation we observe on the Internet today:

> "Behold, they are one people, and they have all one language; and this is only the beginning of what they will do; and nothing that they propose to do will now be impossible for them. Come, let us go down, and there confuse their language, that they may not understand one another's speech."

*Department of Computer Science, University of Chicago, 1100 E. 58th Street, Chicago, IL 60637, Phone: +1-773-702-5680,Fax: +1-773-702-8487,levow@cs.uchicago.edu,

†College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, oard@umd.edu

‡Department of Linguistics and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, resnik@umiacs.umd.edu

§This work was supported in part by DARPA contract N6600197C8540, DARPA cooperative agreement N660010028910, and NSF grant EIA0130422.

Of course, many linguists might dispute this explanation for the diversity evident in human language. Whatever the cause, overcoming the language barrier has been the focus of great interest and substantial investment since the dawn of the computer age. Early efforts proved to be disappointing, in part because the theory, techniques and resources available at the time were not sufficient to automatically produce fluent translation of unrestricted text (ALP, 1966). The situation has improved somewhat in recent years, as new techniques have been developed ((Brown et al., 1990) and successors) and because the emergence of the World Wide Web has provided a strong forcing function. Much of the present focus of application development has been characterized by Church and Hovy as seeking "good applications for crummy machine translation" (Church and Hovy, 1993). Among the uses that have been found, few have been as successful as Cross-Language Information Retrieval (CLIR).

The goal of a CLIR system is to help searchers find documents that are written in languages that are different from the language in which their query is expressed.[1] This can be done by constructing a mapping between the query and document languages, or by mapping both the query and document representations into some third feature space. The first approach is often referred to as "query translation" if done at query time and as "document translation" if done at indexing time, but in practice both approaches require that document-language evidence be used to compute query-language term weights that can then be combined as if the documents had been written in the query language.

In all cases, however, a key element is the mechanism to map between languages. This translation knowledge can be encoded in different forms - as a data structure of query and document language term correspondences in a machine readable dictionary or as an algorithm, such as a machine translation or machine transliteration. While all of these forms are effective, the latter require substantial investment in time and resources for development and thus may not be widely or readily available for many language pairs. Therefore, we focus in this article on the machine-readable dictionary in its simplest form, a bilingual term list. Because of its simplicity, such pairwise lists of translation correspondences are readily available for many language pairs and are relatively easy to construct if unavailable. We identify techniques that allow the CLIR system to best exploit these simple resources and concurrently identify general issues and approaches that have bearing on CLIR techniques that employ more complex encodings of translation knowledge.

In this article, we draw together a body of work that has not previously been accessible in a single source in a way that makes three key contributions. First, we present a holistic view of issues that have previously been presented only in isolation, and often in different communities (e.g., information retrieval and computational

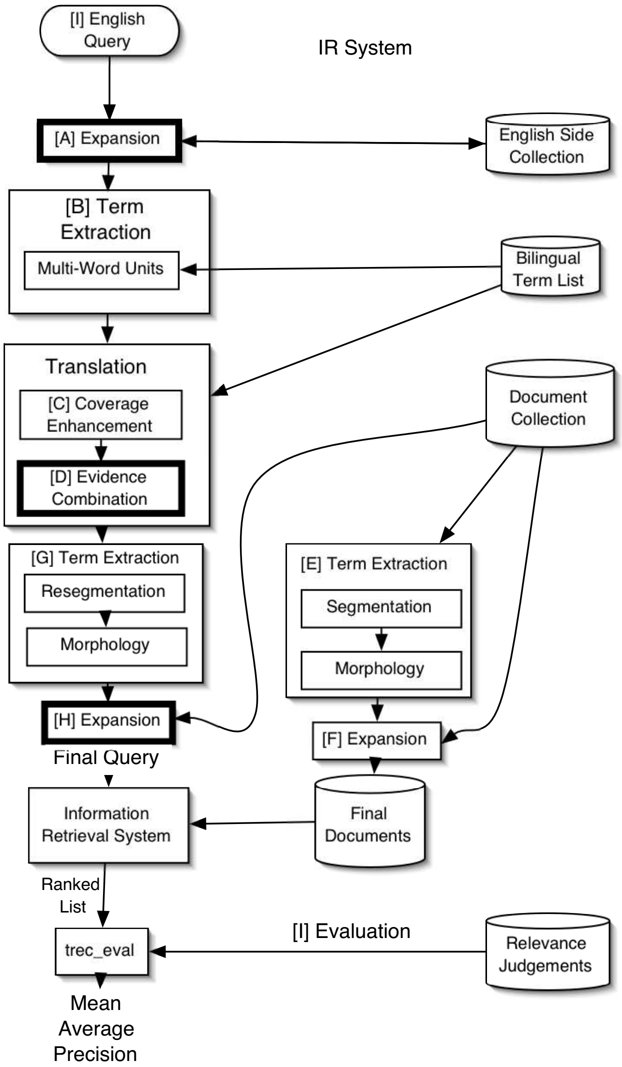
[1] We use the term "document" broadly here to mean any linguistic expression, whether stored as character code in a computer, printed on paper, or spoken. For ease of presentation, we assume that documents in other forms are converted into character codes using optical character recognition or speech recognition prior to indexing, and do not treat those details further in this article.

linguistics). Second, we introduce a unified framework based on mapping evidence about meaning across languages, casting widely used techniques such as balanced and structured translation in that framework as a way of illustrating their relative strengths. And third, we present a comprehensive set of contrastive experiments to illustrate the effects of each technique that we describe, together with new insights based on those results.

Our goal in this article is not merely to describe the state of the art, but to illustrate the effect of the techniques that we describe on retrieval effectiveness for languages with different characteristics. This naturally leads to the question of what system architecture to choose in order to make informative comparisons and what measures to use to make those comparisons. We have chosen a query translation architecture that illustrates the full range of opportunities to improve retrieval effectiveness. Furthermore, as a practical consideration, repeated trials with alternate query translation techniques are more easily run than those with alternate document translation techniques. But our goal in this paper is to present a framework for considering the fundamental issues in dictionary-based CLIR, and those issues will naturally be important considerations in the design of any dictionary-based CLIR system, regardless of the specific architecture adopted. In our experiments, we demonstrate the impact of various techniques on retrieval effectiveness using standard large-scale test collections for languages exhibiting a range of interesting linguistic phenomena. Specifically, we perform experiments using English language queries with document collections in French and Mandarin Chinese for all experimental conditions,[2] and German and Arabic to illustrate some specialized processing.

Figure 1 illustrates the data flow between the key components in our reference architecture. Our dictionary-based query translation architecture consists of two streams of processing, for the query and documents. Close observation will reveal substantial parallelism in the processing of these two streams, as well as symmetry in pre- and post-translation query processing. Specifically, we exploit methods for suitable term extraction and pseudo-relevance feedback expansion at three main points in the retrieval architecture: before document indexing, before query translation, and after query translation. The discussion and experiments throughout the paper will highlight both similarities in the techniques employed at these different stages of processing and differences in the goals and optimization criteria necessary at each stage. Different targets for matching - in the dictionary for pre-translation processing and between the document and translated queries at the other points - influence the specific strategies used as well as the effectiveness of the techniques. The translation process bridges the language gap, and the information retrieval system finally performs the actual query to document match producing a ranked list of documents. Comparison of this ranked list to relevance judgments yields our experimental figure of merit.

[2] We applied document expansion only to French.



	A	B	C	D	E	F	G	H	I
Description section	**3.2**	3.1	4.1	**4.2**	2.1	2.2	4.4	**4.5**	5.1
Results section	**5.6**	5.5	5.9	**5.10**	5.4	5.8	5.4	**5.7**	

Figure 1: CLIR architecture. Letters are keyed to section numbers where components are discussed. Bolding indicates key contributions of this paper.

The remainder of this article is organized as follows.[3] Section 2 describes in some detail document processing for a dictionary-based cross-language setting, introducing the specific techniques used in our experiments. Sections 3 through 4.2 provide a similar level of detail for query processing and for the use of translation knowledge to map between the document and query languages. Experiment results that illustrate the effects of specific techniques appear in section 5.

2 Document Processing

Having discussed our CLIR reference architecture in general terms, we proceed in this section to a discussion of document processing that considers the issues and the alternative methods in greater detail, while also introducing the specific methods used in our later experiments.

In this section we focus on two critical elements in document processing, index term extraction and document expansion, with an emphasis on issues relevant in a cross-language setting. These two steps can be viewed as components in constructing the representation of document content that will be used in retrieval. In the first step, a document is characterized by the set of terms that appear within — though, as we shall see, we may wish to interpret *appear* somewhat indirectly. An English document about crude oil pipelines in Afghanistan can be characterized by terms like *oil*, *pipelines*, and *Afghanistan*; one might also wish to include in the characterization words like *pipeline* and *pipe* that appear implicitly, the better to match terms that could appear in queries seeking documents like this one.

In the second step, one takes this idea of implicitly represented terms further. If the document contains terms like *oil*, *pipelines*, and *Afghanistan*, the concepts underlying the document probably also involve terms like *petroleum*, *gas*, *capacity*, and *kilometer*. Document expansion is the process of adding such terms to the document representation, thereby making explicit those terms that are sufficiently related to the document in a conceptual sense.

2.1 Extraction of Indexing Terms

Index term extraction is relatively straightforward in English, but freely compounding languages such as German and unsegmented languages such as Chinese pose additional challenges. As is well known, the easiest way to extract indexing terms from a document, recognizing tokens separated by whitespace, is often too simple. Here we describe a range of techniques that provide better results given the challenges presented by a representative range of languages including English, French, Arabic, Chinese, and German. These approaches fall into two main categories, automatically segmenting the text stream into a single sequence of non-overlapping words,

[3]Owing to space limitations, we assume that the reader is familiar with the central issues and techniques for ranked retrieval in monolingual applications. See W.B.Frakes and Baeza-Yates (1992) for relevant background.

which might then be subjected to further processing such as stemming, and indexing overlapping character sequences.

2.1.1 English and French: Tokenization, Clitic Splitting, and Stemming

English and French are written with generally space-delimited words, and simple pattern-based approaches to tokenization work well for separating words from punctuation. In both English and French, clitic splitting is employed to separate morphemes connected to a word by an apostrophe. Techniques typically consist of a twofold process of separating the clitic and then expanding it, e.g. *m'aidez* ⇒ *m' aidez* ⇒ *me aidez.*

After tokenization as above, the resulting words are often normalized via either morphological analysis (e.g. (Koskenniemi, 1983)), mapping inflected verbs such as *aidez* to root forms such as *aider*, or, more commonly, by the process of stemming, which typically involves application of a set of rules for removal of prefixes or suffixes or both. The widely used rule-based approach to stemming pioneered by Porter (1980) stems *continua, continuer, continuait, continuera, continuant, continuerait, continuation, continueront, continue, continuez, continué,* and *continuité* to the single stem *continu.*

Notice that unlike morphological normalization, which usually preserves part-of-speech distinctions, stemming freely collapses across parts of speech, e.g. noun *continuité* and verb *continuez* reduce to the same stem, and may freely produce terms that are not actually words in the language. This increases the likelihood of a match when normalized document and query representations are used.

In our experiments on French documents, we first applied a two-stage clitic separation and expansion approach. We then applied a rule-based Porter-style stemmer, freely availble from http://xapian.org, to normalize across morphological variants.

2.1.2 Arabic: Complex Morphology

The morphology of Arabic is far more complex than that of English or French. Adopting a generative view of Arabic morphology, template-based character insertions are used to convert generalized "roots" (such as "*ktb*" [in the standard transliteration], which serves as the base form for many words that have to do with writing) into more specific "stems" (such as "*ktAb*," which means "book"). Prefixes and suffixes can then be added to these stems to form words, and some common modifiers can be adjoined to the beginning or end of a word to form a limited class of compound forms (e.g., "*wktAbAn*, "and two books"). It is often the case that several roots could be used to generate the same Arabic token, so token-level analysis is often highly ambiguous. State of the art techniques such as two-level finite-state morphology (Beesley, 1998) therefore typically generate several possible, but sometimes highly improbable, analyses. Three approaches to this challenge are possible. The first is to do a full analysis and then select the most probable results based on corpus statistics and/or context (e.g., (Darwish, 2002)). A widely used alternative is to instead apply rule-based techniques to remove common prefixes and suffixes (whether from morphology or compounding) to produce something akin to English or French stemming. This approach is typically referred to as "light stemming" in Arabic, since the resulting "stems" sometimes differ from what would be produced by a full linguistic analysis (Aljlayl and Frieder, 2002). A third approach is corpus-based clustering, in which terms found in some other way (e.g., through light stemming) are grouped into classes based on their distributional characteristics (e.g., (De Roeck and Al-Fares, 2000)). We used the first two approaches for the illustrative experiments described in this article that involve Arabic (Darwish, 2002).

In our experiments we formed four kinds of terms:

- Token, in which only white space was stripped.
- Linguistic stems, in which affixes were stripped using the most likely analysis from the Sebawai morphological analyzer (Darwish, 2002).
- Linguistic roots, in which the most likely Sebawai analysis was used to identify the root (e.g. *alkitab* ⇒ *ktb*).
- Lightly stemmed words, in which affixes were automatically stripped using a simple rule-based system (Al-stem). (e.g. *alkitab* ⇒ *kitab*).

2.1.3 Chinese: Word Segmentation

Spoken languages generally lack any explicit marking of the breaks between words, and some written languages exhibit similar characteristics, lacking between-word spaces or other indications of word boundaries. Two approaches to term extraction are possible in such cases: (1) automatic segmentation, and (2) overlapping character n-grams. We have chosen Chinese to illustrate these approaches.

Automatic segmentation techniques typically model the task as selecting a partition on the sequence of characters that corresponds to word boundary positions (although variants that include aspects of stemming or expansion of contractions have also been explored). A wide variety of techniques have been developed, but all can be cast in a framework of model-based optimization. The simplest example is longest substring matching, in which a sentence is traversed from left to right, removing the longest dictionary term that begins at the present position (or a single character, if no dictionary term is found). This corresponds to minimizing the number of characters that are not covered by a term found in the dictionary using a greedy search strategy. The key ideas in this framework are the function to be optimized (in this case, a function of the chosen partition) and the search strategy to be used to explore the space of possible partitions.

Other optimization functions incorporate the degree of fit to hand-segmented training data (for supervised techniques) (Emerson, 2001) or measures of consistency such as minimum description length (for unsupervised techniques). Greedy search strategies are widely used, but dynamic programming (Barras, Geoffrois, Wu, and Liberman, 1998) or exhaustive enumeration (Jin, 1992) are also sometimes employed.

The difficulty of accurate word segmentation for Chinese has led to extensive use of overlapping character n-grams for indexing Chinese. The idea is to eschew the notion of a definitive segmentation altogether, and instead generate all the character n-grams of a fixed width observed in the text. This abandons any semblance of interpretability for the extracted terms, but it does have the advantage of producing term representations that support good matches when they exist, as well as permitting partial matches.

To illustrate with an English word, if a query contains the word *china*, then the trigram terms generated from the query will provide matches against documents containing occurrences not only of *china*, but also *chinese* and *indochina*, since trigrams *chi* and *hin* are shared. This provides the effect of stemming without the necessity of identifying the token boundaries.

The same approach can also be used for languages with white-space delimited tokens, of course. In practice, the choice of n-gram width varies by language, and tends to correlate with the average size of a morphological unit — hence English is best represented using n-grams for n in the vicinity of 5 (Mayfield and McNamee, 1999), and Chinese is best represented with n of 2 (Meng et al., 2001; Wilkinson, 1997).[4]

For the illustrative experiments in this article that involve Chinese, we have experimented with both heuristic longest-match segmentation, using the NMSU segmenter (Jin, 1992) and with terms based on Chinese character bigrams (Section 5.4).

2.1.4 German: Decompounding

Although German uses white-space to separate words, its well known productivity with respect to compound words raises issues of within-word segmentation similar to those of Chinese. German decompounding can therefore be viewed in the same optimization and search framework. We apply a dictionary-based, greedy approach in our experiments, using the German side of our German-English bilingual term list as the segmentation dictionary. Morphological normalization for German terms — either pre- or post-decompounding — can be addressed using the same sorts of normalization approaches discussed above for English and French.

[4] Chinese characters carry semantic content in ways English characters do not, so on average words are approximately two characters long and individual characters often carry some recognizable component of meaning.

2.2 Document Expansion

A quintessential problem in information retrieval is the fact that the same underlying ideas can be represented in many different ways in the observed text. One approach to this problem is to represent underlying concepts as hidden variables in a probabilistic model (Kraaij and Hiemstra, 1998; Ponte and Croft, 1997). Another is to treat term-based representations as incomplete, and to expand them to include terms representative of the underlying concepts that cannot be extracted explicitly from the text itself. This is the basis for widely used query expansion techniques in monolingual information retrieval.

While query expansion is a well-established technique for both monolingual and cross-language information retrieval, document expansion has only recently been applied to these tasks. The document expansion approach was first proposed by Singhal and Pereira (1999) in the context of spoken document retrieval. Since spoken document retrieval involves search of error-prone automatic speech recognition transcriptions, Singhal et al. introduced document expansion as a way of recovering those words that might have been in the original broadcast but that had been misrecognized. Their results showed that correctly recognized terms yield a topically coherent transcript, while the errors tend not to co-occur in comparable documents. Using the document as a query to a comparable collection typically yields documents that contain some related terms that are highly selective; when those terms are added to the document, improved retrieval effectiveness was observed.

The same idea can be applied in CLIR to find words that the author might have used; this can achieve an effect similar to post-translation query expansion. We expanded the original news stories with the most selective terms from related documents. First all documents underwent basic term extraction as described above. Then each document was reformatted as a query in which all terms were weighted equally. We used the full document collection as a comparable collection to be searched for enriching terms. We then selected the top five ranked documents (excluding the original document itself) as sources of expansion terms. Next we chose highly selective terms from these documents by ranking terms in decreasing order based on inverse document frequency. We added an instance of each term to the original document up to one less than the number of expansion documents in which it appeared, until we had approximately doubled the original document length.[5] This process sought to maintain the fidelity of the term frequency component for term weighting. Finally, we indexed the resulting expanded documents.

3 Query Processing

As is the case for document processing, the processing of queries in dictionary-based CLIR depends on extraction of query terms in order to represent the information need, perhaps with query expansion to augment

[5] More efficient implementations of the document expansion procedure are possible, but this simple approach suffices to illustrate the term selection process and its contribution.

that representation with additional related terms. In this section we elaborate on the details of pre-translation query processing; translation issues are then taken up in Sections 4 and 4.2.

3.1 Pre-translation Term Extraction

In Section 2.1, we focused on the extraction of terms from documents in order to provide term-based representations for retrieval. The process of term extraction from queries involves essentially the same linguistic issues and utilizes many of the same techniques. There is, however, a key difference when extracting terms from queries in a query-translation architecture: the terms extracted from queries are going to form the basis for *translation*, not for matching.

Matching techniques (in monolingual applications) typically seek to enhance recall by conflating differences using stemming and segmentation processes, at the expense of some precision. In contrast, for dictionary-based CLIR, a key concern is mitigating the effects of ambiguity, where multiple terms with multiple senses result in an explosion of translation alternatives. We focus on matching the dictionary at the highest level of selectivity to minimize ambiguity, and then applying backoff strategies to enhance coverage only when exact match translations are unavailable.

As an additional tactic to reduce ambiguity, we take advantage of the observation that multi-word expressions rarely have more than one interpretation (e.g., the word "house" in "White House" cannot be translated in the verb sense meaning "to shelter"); translating multi-word expressions as a unit is well known to be helpful for CLIR (Ballesteros and Croft, 1997). Since the task of identifying useful multi-word expressions can be viewed a variant of the segmentation optimization problem, we applied the greedy longest-match technique described in Section 2.1.

3.2 Pre-translation Expansion

Query expansion is a well-established technique in monolingual information retrieval. Very short (2-3 word) queries are common in some applications (e.g., Web search). Expansion can help to compensate for this kind of incomplete specification of the information need. Brevity can yield ambiguity (reducing precision), or may result in omission of terms that are used by authors of the documents that are sought (reducing recall). Query expansion using pseudo-relevance feedback had been shown to partially overcome these difficulties (Buckley, Salton, Allan, and Singhal, 1994). While expansion in general increases the mean retrieval effectiveness, it is also known to concurrently increase variance across queries.

Ballesteros and Croft (1997) evaluated pre-translation and post-translation query expansion in a Spanish-English cross-language information retrieval task and found that combining pre- and post-translation query expansion improved both precision and recall, with pre-translation expansion improving both precision and recall, and post-translation expansion enhancing precision. Mayfield and MacNamee's ablation experiments on the effect of translation resource size on pre- and post-translation query expansion effectiveness demonstrated the dominant role of pre-translation expansion in providing translatable terms (McNamee and Mayfield, 2002). If too few terms are translated, post-translation expansion can provide little improvement. In pre-translation query expansion, our goal is both that of monolingual query expansion — providing additional terms to refine the query and to enhance the probability of matching the terminology chosen by the authors of the document — and providing additional terms to limit the possibility of failing to translate a concept in the query simply because the particular term is not present in the translation lexicon.

We performed the expansion as follows. We constructed the initial query in the normal manner for INQUERY. We then used INQUERY's relevance feedback process to obtain expansion terms based on the 10 highest ranked retrieved documents from the contemporaneous 1994 Los Angeles Times documents, part of the Cross-Language Evaluation Forum (CLEF) (Peters, 2001) 2000 corpus. Experiments with three, five, and ten expansion documents were conducted with minimal difference in resulting retrieval effectiveness. We concatenated the expansion term set to the original query and used the resulting query as the basis for translation.[6]

4 Translation Knowledge & Query Translation

For CLIR, translation knowledge provides the crucial bridge between the user's information need expressed in one language and document concepts expressed the document language. While approaches using off-the-shelf machine translation translation systems alone or in combination with other translation resources have been shown to be effective for CLIR tasks (Gey, Jiang, Chen, and Larson, 1998), they are limited to the relatively small number of language pairs for which such systems exist. Since our goal is to focus on broadly applicable techniques, we focus on the simplest form of a translation lexicon, bilingual term lists, which are already available for many language pairs and can be constructed relatively easily for others. A bilingual term list is an unordered set of query-language/document-language term translation pairs, often with no translation preference or part-of-speech information. In this section, we first describe the bilingual term lists that we used in the CLIR experiments reported below. We next describe a general methodology for backoff translation that enhances dictionary coverage. We then describe two main strategies for integrating translation evidence and managing ambiguity through different term weighting techniques. We conclude by presenting two methods to further enhance matching of the translated query with the document index through term extraction and expansion processes.

[6]In general, query expansion should involve term reweighting as well, both increasing and decreasing; here we adopt the simple strategy of uniform weighting of original and expansion terms.

Translation Resource	# English Terms	# Document Language Terms	Source
English-French	20,100	35,008	http://www.freedict.com
English-Arabic	137,235	179,152	Web word translations
English-Chinese	199,444	395,216	CETA + http://www.ldc.upenn.edu
English-German	99,357	131,273	http://www.quickdic.de

Table 1: Characterization of bilingual term lists

4.1 Bilingual Term Lists & Optimizing Coverage

Bilingual term lists are easily found on the Web, often having been created initially for use with simple online bilingual dictionary programs. However, since these term lists were constructed for diverse purposes and may derive from diverse sources, their ready availability is a great advantage but their possibly eclectic structure also presents challenges. They may vary dramatically along several dimensions, including number of entries, source style, number of multi-word entries, degree of ambiguity, and mix of surface or root form entries. A key challenge for dictionary-based CLIR systems is to develop techniques to most fully exploit these resources while minimizing any negative impact of their more problematic characteristics. A characterization of the bilingual term lists used in our experiments appears in Table 1. In the cases of English paired with French, Chinese[7], and German, we used existing static bilingual dictionary resources. Lacking an English-Arabic resource of similar quality, we constructed an English-Arabic term list by sending each unique word found in a large collection of English news stories to two Internet-accessible English-to-Arabic translation services and merging the results.[8]

Put simply, dictionary coverage is the fraction of the cross-language synonymy relationships that are contained in the bilingual term list. Many languages exhibit a degree of morphological variation that would make it impractical to build a translation lexicon containing every possible variant of each query language word. This is a source of considerable complexity in the design of machine translation systems, but a simple "backoff translation" strategy works well in CLIR applications.

The key insight behind backoff translation is that distinctions between morphological variants of the same word are typically suppressed for retrieval purposes. Translation lexicons that are assembled from fragmentary sources sometimes contain translations for an inflected form but lack translations for the root form of that word. In such cases, the logical approach is to augment the query language side of the translation lexicon with any missing root forms that can be generated from the inflected forms that are present on the query language side of the lexicon. Since any inflected forms on the target-language side are normally conflated prior to retrieval as described in Section 2.1, it is not normally necessary to alter the form of the translations that are associated

[7]The CETA Chinese translation resource is licensed from MRM Corp., Kensington, MD

[8]This bilingual term list can have at most two translations for an English term, and query terms that did not appear in the English collection would have no known translation.

with the inflected form. If multiple inflected forms match a missing root form on the query language side of the lexicon, the associated translations can be merged to produce a single lexicon entry for the root form. In our implementation, we first attempted to match the surface form in the document directly with forms in the term list to maintain precision, only if there was no match, did we compute roots for first the document form and then the lexical entry. As a lightweight approach to morphology, the backoff translation strategy used stemming — automatic affix removal as described in Section 2.1 — to conflate both derivational and inflectional forms. In languages with fairly regular morphology, the use of stemming with backoff translation has worked well (Resnik, Oard, and Levow, 2001).

In Section 2.1.1 we discussed a range of approaches to normalization of morphological variation for languages like English and French. Here we applied Porter's rule-based stemmer to facilitate matching between the surface forms in the queries and the mix of surface and root forms in the English side of the translation resource.

4.2 Weight Mapping

In early work on dictionary-based query translation, queries were typically formed by including all translations for all of the query terms. When used with "bag of terms" retrieval techniques such as the vector space method in which term contributions are treated as independent, this approach can give undue emphasis to query terms that have many translations. This is generally an undesirable trait for a retrieval system, since terms with fewer translations are usually more specific, and more precise results sets are typically obtained when such terms are highly weighted. This *unbalanced* query formulation technique is still often used as a baseline in CLIR experiments, but better techniques are now known.

An obvious improvement is to rebalance the contribution of each term in some way. The key idea, which we call *balanced* translation, is that the weight associated with each translation of a query term can be averaged to compute a weight for that query term (Leek, Schwartz, and Sistra, 2002; Levow and Oard, 2002). Balanced queries formulated in this way can be thought of as estimating the weights for query-language terms as follows:

$$w_j(D_k) = f(TF_j(D_k), DF(D_k), L_j) \tag{1}$$

$$w'_j(Q_i) = \frac{\sum_{\{k|D_k \in T(Q_i)\}} w_j(D_k)}{|\{k|D_k \in T(Q_i)\}|} \tag{2}$$

where $w_j(D_k)$ is the weight that would be computed for term D_k in document j for monolingual retrieval, $TF_j(D_k)$ is the number of occurrences of term D_k in document j, $DF(D_k)$ is the number of documents that contain term D_k, L_j is the number of term occurrences in document j, f is a function that is increasing in TF, decreasing in DF and decreasing in L, $w'_j(Q_i)$ is the weight of query term Q_i for document j, and $T(Q_i)$ is the set of document-language translations for query-language term Q_i.

One potential problem with balanced translation is that typical term weight functions give more weight to rare terms (because they result in more sharply focused result sets). As a result, the average in equation (2) can be dominated by one or more rare terms that earn a high weight simply because they are rare. This would be reasonable if it were known that the rare term was the correct translation, but in the absence of information about translation probabilities it is generally a good idea to suppress rare translations when common alternatives exist.

4.3 Evidence Mapping

Pirkola (1998) introduced an elegant alternative to balanced query translation that will be referred to as *structured* query translation. The key idea is to use translation knowledge to map evidence about meaning (specifically, within-document term frequency, and across-document "document frequency"), and then to combine that evidence to compute term weights in the query language. Formally:

$$TF_j(Q_i) = \sum_{\{k|Q_i \in T(D_k)\}} TF_j(D_k) \tag{3}$$

$$DF(Q_i) = |\bigcup_{\{k|Q_i \in T(D_k)\}} \{d|D_k \in d\}| \tag{4}$$

$$w'_j(Q_i) = f(TF_j(Q_i), DF(Q_i), L_j) \tag{5}$$

where the symbols are as defined above (Kekäläinen and Järvelin, 1998). In the absence of translation probability information, this amounts to assuming that every translation of a query term that is found in a document is actually an instance of that query term. The net effect is to treat query terms that have *any* translation that appears in many documents as a low-weight "common" term.

For the experiments reported in this paper, we used the INQUERY synonym operator, which implements the computation described above at query time. It can be computationally expensive to compute term weights in this way at query time because the postings file must be traversed to compute the union in equation 5, but efficient indexing-time techniques are available for use when query-time efficiency is a concern (Oard and Ertunc, 2002).

4.4 Post-translation Resegmentation

We observed in Section 3.1 that the criteria for term extraction differ depending on whether one is focusing on extracting terms for matching term-based representations, or whether one is focusing on terms to be translated, with the *resulting translations* to be used for matching. For good translation—avoiding unnecessary ambiguity—it is desirable to extract the most specific terms for which it is possible to obtain translations. For good matching, it is desirable to extract terms that are neither too specific nor too general. It is, however, crucial

Language	Morphological Processing	Source
French	Stems by rule-based affix removal	http://xapian.org
Arabic	Linguistic stems by affix removal with probability	Sebawai http://tides.umiacs.umd.edu/
	Linguistic roots by affix removal and pattern	Sebawai http://tides.umiacs.umd.edu/
	"Light-Stems" by rule-based affix removal	Al-stem (see Section 2.1.2)
Chinese	None	
German	Roots by decision tree analysis	http://www.ims.uni-stuttgart.de/projekte/ corplex/TreeTagger/DecisionTreeTagger.html

Table 2: Morphological processing techniques used in experiments.

that query translation and document segmentation produce the *same* terms, otherwise matching cannot occur. It is therefore sometimes necessary to resegment the results of translation to match the set of terms that were indexed (Meng et al., 2001).

For example, consider the case of the query term "bookstore," for which the German translation is *buchhandlung*. If German decompounding resulted in indexing only *buch* ("book") and *handlung* ("store"), applying the same decompounder to the translation results would make it possible to match the appropriate terms. Note that this approach is superior to segmenting "bookstore" in English before translation, since some translations for "store" would not be appropriate in this context.

That approach would not work with structured query translation, which requires TF and DF statistics for complete translations rather than resegmented constituents. A useful approach in that case is to use position-based indexing to obtain the required statistics from instances of the constituents that are adjacent and in order. We used INQUERY's "ordered distance" operator for this purpose, which performs that computation at query time (efficient indexing-time alternatives are also possible). For consistency, we applied the recombination in unbalanced and balanced query translation and to all segmentable terms.

As with post-translation resegmentation, it is necessary to treat the output of the translation process in the same manner as the terms in the document. Post-translation morphological processing yields forms that are more likely to match similarly processed document terms. While there may be some loss of precision by mapping many surface forms to a single root or stem form, the increase in recall due to this conflation more than outweighs this small change in precision. Thus we apply the techniques described in Section 2.1; the specific implementations appear in Table 2.

4.5 Post-translation expansion

Post-translation query expansion can overcome translation gaps by identifying related terms in the document language to be added to the translated query. In addition to recovering translations of common terms, post-translation query expansion has great potential for recovering named entities in the document language which are often not found in standard translation resources.

We used the translated queries, possibly after pre-translation expansion, as the basis for expansion. We enriched the translated query with a set of expansion terms similarly to the pre-translation case in Section 3.2. We used the document collection to be searched as the source of expansion documents.

5 Experiments

In this section we describe the quantitative results of our experiments applying different strategies to the three main questions of term extraction, expansion, and using translation evidence in turn, for each of the applicable main processing stages, such as document processing, pre-translation query processing, and post-translation query processing (Figure 1).

5.1 Test Collections

We performed experiments using English language queries with document collections in French and Mandarin Chinese for all experimental conditions, and German and Arabic to illustrate some specialized processing. We used title and title+description queries (with no stop structure removal) to characterize the information need. Title format yields short (usually 1-3 word) queries that are typical of what a Web searcher would provide, for example, "Architecture in Berlin". Title+description queries augment the brief title with a few explanatory sentence similar to those a searcher might use when discussing their information need with a trained search intermediary. We used the test collections described below:

French: We utilize the Cross-language Evaluation Forum (CLEF) 2000 French subset, including 72,000 documents from the French newspaper *Le Monde* in 1994 with the 40 English queries.

Arabic: We utilize the Arabic test collection from the Text Retrieval Conference (TREC-2002) CLIR track, which includes 383,872 *Agence France Presse* news stories and 50 English queries (Oard and Gey, 2002).

Chinese: We utilize the TREC-5 Chinese monolingual track, including 164,600 documents from the *People's Daily* 1991-1993 and *New China News Agency (Xinhua)* 1994-1995 with 28 queries manually translated into English from their original Chinese form.

German: We utilize the Cross-language Evaluation Forum (CLEF) 2000 German subset, including 153,694 documents from *Der Spiegel* and *Frankfurter Rundschau* 1994-1995 with the same 40 English CLEF 2000 queries used for the French above.

For the experiments reported below, we used the INQUERY information retrieval system (version 3.1p1) to produce retrieval status values for the documents in the test collections. The documents were then sorted in order of decreasing retrieval status value to form a list in an order that approximates a decreasing degree of relevance to the searcher's query. We then used the standard trec_eval software to compute uninterpolated mean average precision based on the pooled relevance judgments released by the corresponding evaluations. We adopted the convention that values of $p < 0.05$ for a Wilcoxon signed ranks test on a pair of retrieval results is considered significant. For the tabular presentation of contrastive CLIR results below, the odd columns of each condition generally represent the base condition and the even columns the contrasting test condition. A table describing each experimental condition, identifying the values for the processing parameters held constant as well as the primary contrast, accompanies each set of results. The condition table above each bar chart of results should be interpreted as follows. The first row identifies the processing stage, i.e. document or pre-translation query processing. The second row specifies the actions performed at that stage of processing, e.g. term extraction, such as segmentation and morphological processing, and expansion in the document processing stage. The cells contain the settings for each of the above actions that apply throughout the experiment. **Varies** appears in the cells for the primary contrasts illustrated in the table. Where the difference between the two conditions reaches statistical significance, the mean average precision is given in bold text.

5.2 Baseline Runs

Retrieval effectiveness measures are well known to be sensitive to collection characteristics, so we focus on comparisons of different techniques on the same collection. In this section, we establish monolingual baselines that serve as a reference point for assessing the relative effectiveness of our cross-language techniques. Each of the evaluation collections includes manually constructed monolingual queries. In each case, we applied standard term extraction techniques for the monolingual case: stemming for French, decompounding for German, overlapping bigram formulation for Chinese, and light stemming for Arabic. We used the same document processing, information retrieval engine, and evaluation methodology as in the cross-language case. For comparison, we included cross-language results for similar query processing and also the best CLIR results obtained under any processing configuration in our current experiment set, in terms of mean average precision and percentage of monolingual effectiveness (Table 3).

These figures should be assessed with the understanding that all CLIR results incorporate multiple translations and thus implicitly involve some query expansion effects. Due to both explicit expansion techniques and these implicit expansion effects, CLIR effectiveness may often exceed a simple monolingual baseline, even though most straightforward CLIR approaches typically achieve 50% of monolingual. These results thus illustrate a creditable performance with simple processing and substantial improvements with more sophisticated techniques.

Condition	Monolingual	CLIR	% Mono	CLIR Best	% Mono
French Stem TD	0.3832	0.2311	60.3%	0.3288	85.8%
German Decomp TD	0.2937	0.2122	72.3%	0.2122	72.3%
Chinese Bigram TD	0.3299	0.2832	85.8%	0.3349	101.5%
Arabic Lightstem TD	0.286	0.197	68.9%	0.197	68.9%

Table 3: Contrasting retrieval effectiveness within and across languages

5.3 Language and Dictionary Effects

In the following sections, we illustrate the relative effectiveness of different experimental conditions in the cross-language context. We demonstrate that many of the differences observed depend on two main factors: the languages involved, both individual language characteristics and interactions between the particular cross-language pair of interest, and the characteristics of the translation resource.

Although most of the techniques discussed in this article are, in principle, language independent, their relative applicability and impact on retrieval effectiveness depends on the languages involved. These effects derive from the different characteristics of the query and document languages and the interactions between the two. For example. one would expect morphological processing to have greater impact on languages with richer morphology.

In dictionary-based CLIR, the translation resource naturally plays a pivotal role. The results in several of our experiments demonstrate effects of differences in dictionary composition. The dictionaries used in these experiments vary dramatically along several dimensions: 1) size: from 40,000 to almost 400,000 entries, 2) source style: from primarily human readable to primarily machine readable, 3) form of entries: from none to 25% multi-word English entries, 4) number of translation alternatives: from typically 2 or 3 to as many as 50, 5) effective translation ambiguity: query by-token fan-out, ranging from 2 to 10 translations, and 6) form of English and foreign language entries: from primarily root forms to a mix of root and surface forms to almost exclusively surface forms.

5.4 Cross-language Runs: Term Extraction in Document and Post-translation Query Processing

The goals of term extraction differ across phases of processing. Crucially, the terms extracted in document processing and in post-translation query processing must be consistent to enable matching in the retrieval phase. In pre-translation query processing, the term extraction process should enhance translation through consistency with the form of the bilingual term list.

We consider two forms of term extraction to enhance matching of concepts between the document and translated query: sub-word segmentation for German and Chinese and morphological analysis for French and

Arabic.

5.4.1 Sub-word Segmentation

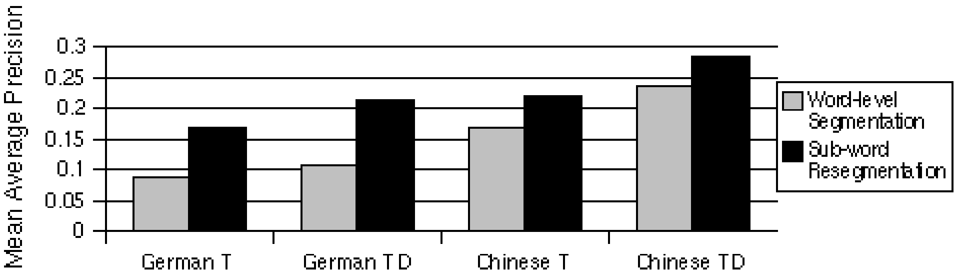
We applied a dictionary-based greedy approach to German decompounding and bigram segmentation to Chinese. We contrast these sub-word approaches with white-space delimited words for German and NMSU's word-based segmentation for Chinese (labeled *Words* in the condition table) (Figure 2).

We found significant improvements in retrieval effectiveness for the new term segmentation processes for both German (T, $p \leq 0.01$ and TD, $p \leq 0.002$) and Chinese(TD, $p \leq 0.025$). These results are supported by prior experiments in sub-word segmentation in Chinese, including (Meng et al., 2001; Nie, Gao, Zhang, and Zhou, 2000; Wilkinson, 1997) where overlapping character n-grams, typically bigrams, improved retrieval effectiveness for Chinese mono- and cross-language contexts over a single word-level segmentation using rule-based or probabilistic approaches. Likewise, the effectiveness of our simple greedy heuristic for German decompounding added further to the work demonstrating the utility of decompounding (Braschler & Schauble, 2001; Chen, 2002; Mayfield and McNamee, 1999). Braschler and Schauble (2001) applied a rule-based decompounder for German nouns, while Chen (2002) used a probabilistic strategy to select decompositions. Similarly to the Chinese case, Mayfield and McNamee (1999) applied character 5-gram segmentation to perform pseudo-decompounding. Each of these sub-word segmentation strategies, while using a variety of optimization strategies for unit selection, yields units which substantially enhance recall, offsetting any loss of precision, by matching semantically meaningful units below the word level. Meng et al. (2001) has suggested additional improvements for appropriately tuned combinations of word and sub-word units; a general approach to such integration is an avenue for future research.

5.4.2 Morphology

We applied simple rule-based stemming for French, contrasted with the unstemmed case (Figure 3). While the impact on retrieval effectiveness did not reach significance in this case, the French queries achieved more than 10% relative increase in mean average precision (mAP). Stemming is commonly used in monolingual retrieval for efficient indexing, though it rarely yields large improvements for languages with simpler morphology (e.g. English) when documents are sufficiently long (Krovetz, 1993). In the case of more morphologically complex languages such as Arabic (see below), such analysis is essential even in the monolingual case. In the cross-language case, due to the mix of root and surface forms present in the translation resource, stemming can play a more significant normalizing role. Oard, Levow, and Cabezas (2001) observed a significant jump in retrieval effectiveness when stemming was linked with backoff translation. The current experiments echo those

Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Varies No Stemming	No	Multi-word Units	No	Structured	Stemming	**Varies** No Stemming	No



	German		Chinese	
Term Extraction	Word-level Segmentation	Sub-word Resegmentation	Word-level Segmentation	Sub-word Resegmentation
T	0.0882	**0.1694**	0.1685	**0.2206**
TD	0.1089	**0.2122**	0.2362	**0.2832**

Figure 2: Retrieval effectiveness with and without post-translation resegmentation using sub-word units in German and Chinese.

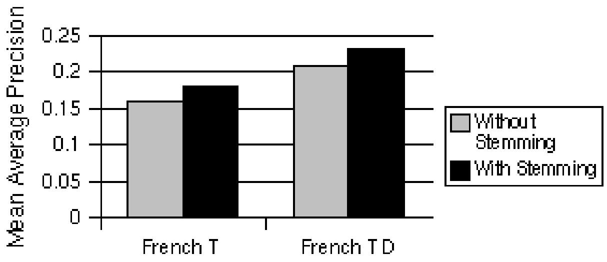
of Sheridan, Wechsler, and Schäuble (1997), where morphological analysis enhanced efficiency and effectiveness in the English-French case.

For Arabic (Figure 4) we contrast four forms for matching: whitespace delimited tokens, linguistic stems, linguistic roots, and lightly stemmed words. We found that any form of normalization significantly enhanced retrieval effectiveness ($p \leq 0.01$) and that light stemming was very effective. Arabic, with more complex morphology, benefits particularly dramatically from suitable morphological handling. Furthermore, interestingly, the same underlying technique of simple affix removal can be applied across a wide range of languages and is effective even for the most complex. These results are consistent with prior work finding significant improvements for Arabic CLIR using a wide range of morphological processing techniques with light stemming yielding improvements comparable to more complex techniques (Larkey, Ballesteros, and Connell, 2002; Aljlayl and Frieder, 2002; Darwish, 2002). The findings apply to dictionary-based techniques employing manually constructed bilingual term lists with an eclectic mix of surface and root forms, contrasting with the term list used by Xu, Fraser, and Weischedel (2002) derived from parallel corpora, that yields better coverage of surface forms and thus exhibits less impact for morphological processing.

5.5 Cross-language Runs: Term Extraction: Pre-translation Query Processing

In general, multi-word units are less ambiguous than single words. We compared translation with space-delimited units to using multi-word units in Figure 5. We found that translation with multi-word units, derived

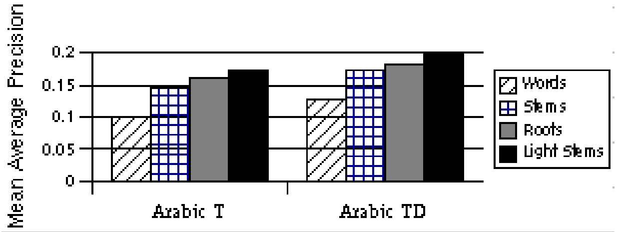
Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Words **Varies**	No	Multi-word Units	No	Structured	Stemming	Words **Varies**	No



	French	
Morphology	Without Stemming	With Stemming
T	0.1596	0.1804
TD	0.2087	0.2311

Figure 3: Retrieval effectiveness with and without document and post-translation query morphological analysis in French.

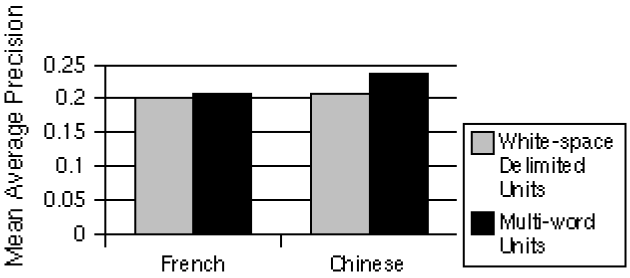
Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Words **Varies**	No	White-space Delimited Units	No	Structured	No	Words **Varies**	No



Morphology	Words	Stems	Roots	Light Stems
Arabic T	0.101	**0.1455**	**0.1599**	**0.1735**
Arabic TD	0.1259	**0.1718**	**0.1814**	**0.1970**

Figure 4: Retrieval effectiveness for alternative ways of normalizing Arabic terms.

Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Words No Stemming	No	**Varies**	No	Structured	Stemming	Words No Stemming	No



French		Chinese	
White-space Delimited Units	Multi-word Units	White-space Delimited Units	Multi-word Units
0.2017	0.2087	0.2085	0.2362

Figure 5: Retrieval effectiveness with and without pre-translation query term extraction of multi-word units.

by greedy segmentation of the queries, yields no significant improvement over translation with white-space delimited units. However, the Chinese queries achieved a 10% relative increase in mAP, while French showed little effect. If this apparent contrast represents a real effect, it suggests the impact of differences in translation resources. Previous work on multi-word (or phrasal) translation (Ballesteros and Croft, 1997; Meng et al., 2001) has demonstrated substantial improvements, where such units are available. Since we only translated multi-word units attested in the English side of the bilingual term list, we expected to see a direct impact of the number and quality of multi-word dictionary entries. Approximately 10% of entries in the French term list consist of multi-word English forms, with only 5% of entries being multi-word units more interesting than "the X". On the other hand, since the English-Chinese term list was built by inverting a manually constructed Chinese-English lexicon designed for human use, almost 25% of the English headwords are multi-word. Short title-only queries did not benefit at all, as no such units were identified. In the case of the title+description queries, only six multi-word terms were identified for translation into French, of which three were of the form "the X", while 36 multi-word units were identified in the queries (some repeatedly) to be translated to Chinese.

5.6 Cross-language Runs: Expansion: Pre-translation Query Processing

We extended the pseudo-relevance feedback expansion approach well-established for monolingual information retrieval to the cross-language domain. We applied this expansion technique to the three possible points in the cross-language retrieval query translation architecture: during query processing before translation, during query processing after translation, and during document processing.

We applied pre-translation query expansion to enrich the description of the searcher's information need. For most French TD queries, we obtained statistically significant improvement in cross-language retrieval effectiveness (Figure 6) . Pre-translation query expansion was actually detrimental for Mandarin Chinese. These results contrast sharply with standard results for pseudo-relevance feedback query expansion, monolingually (Buckley et al., 1994) and across languages (Ballesteros and Croft, 1997) which found improvements for both precision and recall. The strongest contrast is with McNamee and Mayfield (2002) where an ablation study with progressively impoverished translation resources demonstrated near 40% monolingual retrieval effectiveness in the English-French case using pre-translation expansion, with an empty translation resource. These findings were used to argue for the primary importance of pre-translation query expansion, since translation relies on the presence of some translatable terms, made more likely by expansion, and cognate matching alone on original and expansion terms yielded modest effectiveness. While we see comparable significant improvements for the French case in our experiments, how can we explain the contrast for Chinese?

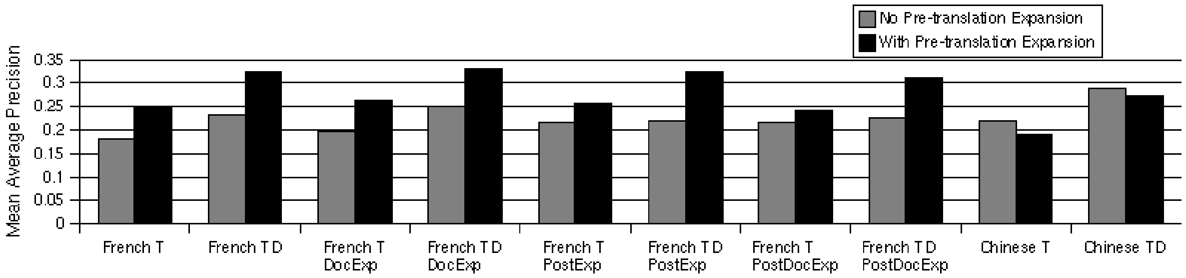
This contrast can be explained simply by the interaction between query and document language. When expanding the original English query with English terms, potentially useful enriching terms are introduced regardless of document language. However, with French documents, both translatable terms and untranslatable cognates, which often include particularly selective named entities, can match in the retrieval process. In contrast, since English and Chinese do not share a common orthography, only translatable terms can enhance retrieval, and named entities are relatively unlikely to be present in the translation resource. As a result, pre-translation query expansion can provide significantly greater benefit in language pairs of greater similarity, such as those with a common orthography.

We utilized the 1994 *LA Times* collection for expansion. These documents are general newswire sources that overlap the time epoch of both the French and Chinese. International topics of interest are better covered than domestic issues for either the French or Chinese collection. Some of the CLEF topics are known to be covered by the English collection. However, even with paired English-Chinese collections designed for shared topics such as the Topic Detection and Tracking (TDT) collections, pre-translation expansion is limited in utility (Levow, 2003).

5.7 Cross-language Runs: Expansion: Post-translation Query Processing

Post-translation query expansion likewise adds terms to the query to improve retrieval effectiveness, for example, to recover from translation gaps or to enrich the terminology in the translated query beyond that available in the bilingual term list. We found significant improvements in retrieval effectiveness through post-translation query expansion in Mandarin Chinese (Figure 7). For French, although there was a 10% relative increase in mAP for title-only queries, no significant improvement was obtained and pre-translation query

Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Sub-word Units Stemming	**Varies**	Multi-word Units	**Varies**	Structured	Stemming	Sub-Word Units Stemming	**Varies**



Query Form	French								Chinese	
	Base Query		DocExp		PostExp		PostDocExp		Base Query	
	1	2	3	4	5	6	7	8	9	10
T	0.1804	0.2469	0.1959	0.2636	0.222	0.2561	0.2164	0.2405	0.2206	0.192
TD	0.2311	0.3226	0.2495	**0.3288**	0.2204	**0.3248**	0.2264	**0.3103**	0.2894	0.2728

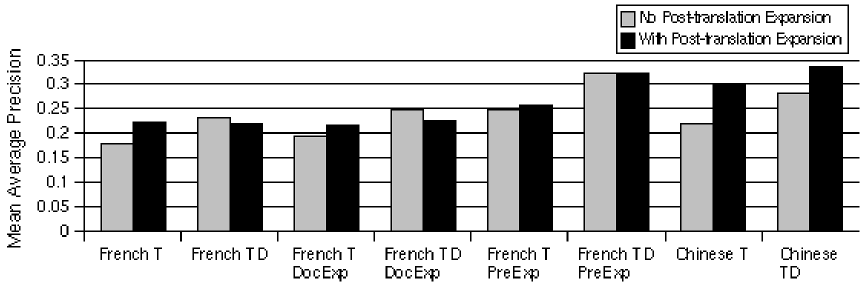
Figure 6: Retrieval effectiveness with and without pre-translation query expansion with and without other expansion. Even numbered columns include pre-translation expansion, odd numbered columns do not.

expansion generally outperformed post-translation query expansion.

Here again, the results for post-translation query expansion in French support work by Ballesteros and Croft (1997), McNamee and Mayfield (2002). In particular, as argued by McNamee and Mayfield (2002) for European language pairs, pre-translation query expansion can have a greater impact than post-translation query expansion. However, we need to explain the dramatic reversal in effectiveness and improvement for Chinese in the post-translation case, both relative to pre-translation and relative to the French case. We believe that the significant improvement for post-translation query expansion for Chinese can be attributed to the recovery of untranslatable terms by the expansion process. Many useful terms, such as named entities, are absent from translation resources and thus may be lost whether introduced in the original query or in expansion. In contrast, such terms are readily available in the Chinese language expansion documents and can enrich the search. In a comparable study of pre- and post-translation expansion on the document side in Chinese-English CLIR, Levow (2003) found a similar trend of dramatically greater utility for post-translation expansion for languages with different orthographies. While the apparent increase in mAP suggests that French benefits from post-translation query expansion, the possibility of orthographic matching of untranslatable cognates enhances both baseline, unexpanded retrieval and pre-translation expansion. Post-translation expansion allows the English-Chinese case to overcome the handicap of the absence of orthographic cognates.

We further note that while some significant improvement was found for French in combining pre-translation

Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Sub-word Units Stemming	**Varies**	Multi-word Units	**Varies**	Structured	Stemming	Sub-Word Units Stemming	**Varies**



Query Form	French						Chinese	
	Base Query		PreTransExp		DocExp		Base Query	
	1	2	3	4	5	6	7	8
T	0.1804	0.2220	0.1959	0.2164	0.2524	0.2560	0.2206	**0.3021**
TD	0.2311	0.2204	0.2495	0.2264	0.3226	0.3248	0.2832	**0.3349**

Figure 7: Retrieval effectiveness with and without post-translation query expansion in French and Chinese. Even numbered columns include post-translation expansion, odd numbered columns do not.

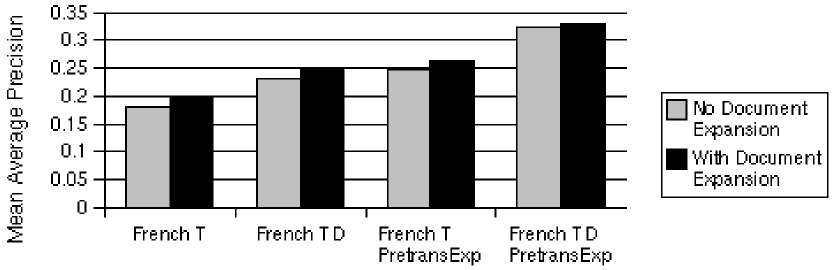
expansion and document expansion, no such improvements are obtained by combining post-translation query expansion and document expansion. We speculate that this phenomenon arises from the fact that we are attempting to exploit the same information (terms in the same document language corpus) albeit from different bases and are faced with diminishing returns.

5.8 Cross-language Runs: Expansion: Document Processing

We performed document expansion as outlined by Singhal and Pereira (1999). We found small increases in mean average precision but no significant improvements in retrieval effectiveness, both with and without pre-translation expansion (Figure 8). One can attribute the limitation of the effectiveness of this technique to the fact that due to the shared orthography of the query and document languages, most terms will translate or match as cognates. Thus the enrichment of the documents with additional terms has little additional improvement.

For contrast, we consider related work in document expansion in CLIR. Lo, Li, Levow, Wang, and Meng (2003) found significant improvements for hybrid word and sub-word document expansion strategies for English-Chinese CLIR. Furthermore, Levow (2003) found a similar trend. Specifically, post-translation expansion yielded significant improvements over the unexpanded case, as well as over the pre-translation expansion case, in a document translation architecture.

Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Sub-word Units Stemming	**Varies**	Multi-word Units	**Varies**	Structured	Stemming	Sub-Word Units Stemming	**Varies**



	Base Query		Pre-translation Expansion	
Expand	No Document Expansion	With Document Expansion	No Document Expansion	With Document Expansion
French T	0.1804	0.1959	0.2524	0.2636
French TD	0.2311	0.2495	0.3226	0.3288

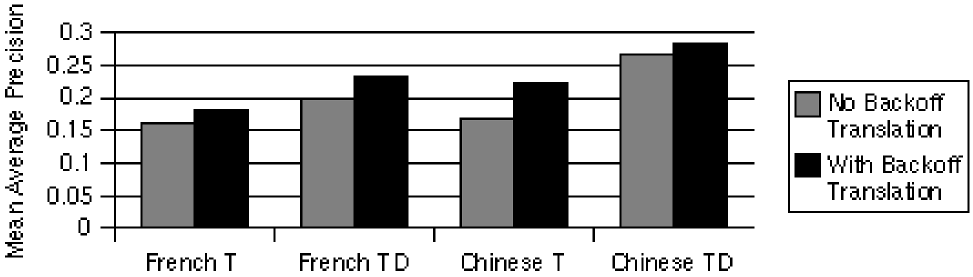
Figure 8: Retrieval effectiveness with and without document expansion in French.

5.9 Cross-language Runs: Backoff Translation

We considered two sets of approaches for handling translation evidence from bilingual term lists, by combining translation evidence from root and surface forms in the query and bilingual term list and by reweighting alternative translations in the term list suitably.

Backoff translation aims to improve recall with minimal effects on precision by combining evidence from root and surface forms. We found more than 10% relative increase in mean average precision across several experiments in French and Chinese, as illustrated in Figure 9, though only for Chinese title queries did these differences reach significance. This modest improvement derives from the ability to match a full variety of surface forms even when the translation resource has a mix of surface and root forms on the English side. Prior work (Oard et al., 2001) demonstrated statistically significant improvements for backoff translation to enhance coverage in a document translation architecture for French and German. Furthermore, Resnik et al. (2001) found further significant improvements by incorporating a noisy translation resource induced from a parallel corpus as a final stage backoff resource, where a direct merge yielded no improvement. Thus backoff translation is a flexible method to enhance coverage using translation resources of differing characteristics. Furthermore, the different levels of impact suggest that greater morphological variation of the source language increases the utility of backoff, as the French and German cases above were more greatly affected than the less heavily inflected English queries in the current experiments.

Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Sub-word Units Stemming	No	Multi-word Units	No	Structured	**Varies**	Sub-Word Units Stemming	No



	French		Chinese	
	No Backoff Translation	With Backoff Translation	No Backoff Translation	With Backoff Translation
T	0.1610	0.1804	0.1681	**0.2206**
TD	0.1968	0.2311	0.2664	0.2832

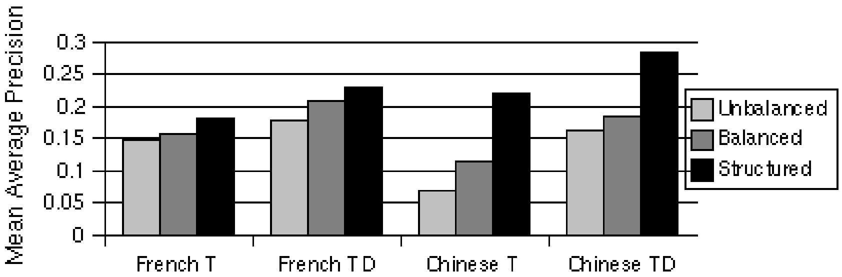
Figure 9: Retrieval effectiveness with and without backoff translation to increase translation coverage in French and Chinese.

5.10 Cross-language Runs: Translation Evidence Combination

We compared three main strategies in query translation to combine evidence from alternative translations: unbalanced translation where alternatives are combined with their original weights, balanced translation where each translation receives a weight mapped to the average of the document-based weights of the alternatives, and structured query translation where the weights are assigned by treating all translations of a query term that appear in the document as evidence for the term. We found that structured query translation outperforms all other methods of combining translation alternatives in each of the languages studied. The improvement was highly significant ($p \leq 0.001$) for Chinese (Figure 10). Structured translation combination significantly outperformed not only the naive unbalanced combination strategy, but also the more sophisticated balanced combination technique for the Chinese case.

Both balanced (Levow and Oard, 2002; Leek et al., 2002) and structured (Ballesteros and Croft, 1996; Pirkola, 1998) evidence combination strategies have been shown to outperform unbalanced strategies in a variety of contexts (Meng et al., 2001; Levow and Oard, 2002; Oard et al., 2001). However, the relative impact of the strategies can vary. The current set of experiments provides some additional insight into the factors affecting these strategies. While in the French case, the increase in effectiveness from unbalanced to structured was a moderate 21-30% relative improvement, in the Chinese experiments, the improvement was highly significant, a 79-220% relative improvement. The wide range in number and applicability of possible translations influences

Document		Pre-translation		Translation		Post-translation	
Extraction	Expand	Extraction	Expand	Query Trans.	Backoff Trans.	Extraction	Expand
Sub-word Units Stemming	No	Multi-word Units	No	**Varies**	Stemming	Sub-Word Units Stemming	No



Query Form	French			Chinese		
	Unbalanced	Balanced	Structured	Unbalanced	Balanced	Structured
T	0.1488	0.1560	0.1804	0.0688	0.1145	**0.2206**
TD	0.1769	0.2072	**0.2311**	0.1594	0.1855	**0.2832**

Figure 10: Retrieval effectiveness for structured query formulation over unbalanced and balanced translation evidence combination strategies.

these results. In the French term list, most terms have close to the average 1.3 translations and are terms in common use. Thus, the deleterious effects of unbalanced translation are less evident. In the English-Chinese term list, the average number of translations is only 1.9, but some terms have as many as 50 translations, and some of the translations are special-purpose. Furthermore, if one considers the true effective degree of translation ambiguity by comparing the number of translated words in the query form to the number of words output in the document language, we find that the queries translated to French approximately double in length while the queries translated to Chinese are approximately ten times their untranslated length. As a result, the effects of appropriately combining translation evidence were magnified in the highly ambiguous Chinese case relative to the less productive French case. Structured translation combination moderated the effect of individual high IDF alternatives, which have a dramatic effect in unbalanced translation.

6 Conclusion

Dictionary-based CLIR techniques have much in common with both machine translation and information retrieval. The dependence on language-specific morphological analysis and the use of context (e.g., multi-word expressions) to constrain translation ambiguity are both evocative of machine translation techniques. The tolerance for ambiguity and incomplete expression are features that are also key to the utility of “bag-of-terms” monolingual retrieval systems. But we have seen that dictionary-based CLIR techniques also introduce new issues, including asymmetry in term selection for translation and matching, additional expansion opportunities, and a focus on nuanced mappings of different types of evidence about meaning across languages. The resulting picture is more complex than a simple modular combination of translation and retrieval, and–when well conceived–that complexity is rewarded by enhanced retrieval effectiveness.

Our choice of query translation as a reference architecture was not meant to imply that all the operations we have considered should be implemented at query time; indeed, we have pointed out several cases in which indexing-time implementations can result in enhanced query-time efficiency. But casting the issues in a consistent framework helps to highlight cases where techniques have differential effect across languages. For example, we have shown that expansion effects are sensitive to the presence of orthographic cognates. We also saw that document expansion and post-translation query expansion produced no further gain when used together, indicating that the choice between the two approaches can be based on implementation considerations.

We also gained some insights from casting into a unified framework a range of techniques that have previously been explored separately. For example, a unified view of term selection reveals the tension between competing optimization criteria for pre-translation and document-language term selection that exists when using structured queries to perform evidence mapping. As a second example, dictionary fanout is often averaged over all source-language terms in a dictionary; explaining CLIR results in our evidence-translation framework led us to focus on the fanout across the query terms—for which the fanout was greater by a factor of five! As a third example, viewing segmentation, decompounding and even phrase identification as a model-based optimization task helps to see commonalities among related techniques that are applied in differing situations.

A focus on dictionary-based CLIR is useful because many of the techniques are now well understood, so the remaining interactions can be well characterized. But dictionaries are only one source of translation knowledge. Two other knowledge sources are of particular importance to CLIR: corpus statistics, and human expertise. Lessons learned with dictionary-based techniques can prove useful for those tasks as well. For example, recent research in both areas makes use of carefully considered pre-translation and document-language term selection, and we have recently extended the idea of backoff translation to include a request from the user for an alternate query term when no suitable translation is known to the system. Other ideas from dictionary-based CLIR might also find productive application with corpus-based or interactive techniques; for example, extending evidence translation to leverage corpus-based translation probabilities seems like a promising direction for future work. And, of course, corpora and people are natural sources to which to look when seeking to enhance a translation dictionary. The full story on CLIR may not be written for some time to come, but there is much to be learned from the way in which the simplest representation of translation knowledge, a term list extracted form a bilingual dictionary, can be used to unlock access to information in other languages.

Acknowledgments

The authors wish to thank Clara Cabezas, Kareem Darwish, Dina Demner-Fushman, and Funda Ertunc for their assistance with the experiments reported above.

References

(1966). Language and machines: Computers in translation and linguistics..

Aljlayl, M., & Frieder, O. (2002). On Arabic search: improving the retrieval effectiveness via a light stemming approach. In *Proceedings of Conference on Information and Knowledge Management CIKM 2002*, pp. 340–347.

Ballesteros, L., & Croft, B. (1996). Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems*, pp. 791–801. http://ciir.cs.umass.edu/info/psfiles/irpubs/ir.html.

Ballesteros, L., & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 84–91.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. In *Proceedings of the First International Conference on Language Resources & Evaluation (LREC)*, pp. 1373–1376.

Beesley, K. R. (1998). Arabic morphological analysis on the internet. In *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing.*

Braschler, M., and Schauble, P. (2001). Expertiments with the Eurospider retrieval system for CLEF2000. In *Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 2000, Revised Papers*, Vol. LNCS 2069 of *Lecture Notes in Computer Science*, pp. 140–148 Heidelberg. Springer-Verlag.

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

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1994). Automatic query expansion using SMART: TREC 3. In Harman, D. K. (Ed.), *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 69–80. NIST. NIST Special Publication 500-225.

Chen, A. (2002). Cross-language retrieval experiments at CLEF 2002. In Peters, C. (Ed.), *Working Notes for the CLEF 2002 Workshop*. http://clef.iei.pi.cnr.it:2002/workshop2002/WN/01.pdf.

Church, K. W., & Hovy, E. H. (1993). Good applications for crummy machine translation. *Machine Translation*, *8*, 239–258.

Darwish, K. (2002). Building a shallow morphological analyzer in one day. In *Proceedings of ACL Workshop on Computational Approaches to Semitic Languages*, pp. 47–54.

De Roeck, A., & Al-Fares, W. (2000). A morphologically sensitive clustering algorithm for identifying Arabic roots. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 199–206 Hong Kong.

Emerson, T. (2001). Segmentation of Chinese text. *MultiLingual Computing & Technology*, *12*(2), 49–52.

Gey, F. C., Jiang, H., Chen, A., & Larson, R. R. (1998). Manual queries and machine translation in cross-language retrieval and interactive retrieval with Cheshire II at TREC-7. In Voorhees, E., & Harman, D. (Eds.), *The Seventh Text REtrieval Conference (TREC-7)*, pp. 527–540. NIST. NIST Special Publication 500-242.

Jin, W. (1992). A case study: Chinese segmentation and its disambiguation. Tech. rep. MCCS-92-227, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Kekäläinen, J., & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 130–137.

Koskenniemi, K. (1983). Two-Level Model for Morphological Analysis. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pp. 683–685.

Kraaij, W., & Hiemstra, D. (1998). Cross language retrieval with the TwentyOne system. In Voorhees, E., & Harman, D. K. (Eds.), *The Sixth Text Retrieval Conference (TREC-6)*, pp. 753–761. NIST. NIST Special Publication 500-240.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–202.

Larkey, L., Ballesteros, L., & Connell, M. (2002). Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–282.

Leek, T., Schwartz, R., & Sistra, S. (2002). Probabilistic approaches to topic detection and tracking. In *Topic Detection and Tracking: Event-Based Information Organization*, pp. 67–84. Kluwer, Boston.

Levow, G.-A. (2003). Issues in pre- and post-translation document expansion: Untranslatable cognates and missegmented words. In *Proceedings of Sixth International Workshop on Information Retrieval with Asian Languages*, pp. 77–83.

Levow, G.-A., & Oard, D. W. (2002). Signal boosting for translingual topic tracking: Document expansion and n-best translation. In *Topic Detection and Tracking Research*, pp. 175–196. Kluwer, Boston.

Lo, W.-K., Li, Y.-C., Levow, G.-A., Wang, H.-M., & Meng, H. (2003). Multi-scale document expansion in English-Mandarin cross-language spoken document retrieval. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech2003)*, pp. 2337–2340.

Mayfield, J., & McNamee, P. (1999). Indexing using both n-grams and words. In Voorhees, E., & Harman, D. (Eds.), *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pp. 419–424. NIST Special Publication 500-242.

McNamee, P., & Mayfield, J. (2002). Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 159–166.

Meng, H., Chen, B., Grams, E., Lo, W.-K., Levow, G.-A., Oard, D., Schone, P., Tang, K., & Wang, J. Q. (2001). Mandarin-English Information (MEI): Investigating translingual speech retrieval. In *Proceedings of the First Human Language Technology Conference (HLT)-2001*, pp. 239–245.

Nie, J.-Y., Gao, J., Zhang, J., & Zhou, M. (2000). On the use of words and n-grams for chinese information retrieval. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, pp. 141–148.

Oard, D. W., & Ertunc, F. (2002). Translation-based indexing for cross-language retrieval. In *Advances in Information Retrieval*, Vol. 2291 of *Lecture Notes in Computer Science*, pp. 324–333 Berlin. Springer-Verlag.

Oard, D. W., & Gey, F. C. (2002). The TREC-2002 Arabic-English CLIR track. In Voorhees, E., & Buckland, L. P. (Eds.), *The Eleventh Text Retrieval Conference (TREC-2002).* NIST. NIST Special Publication 500-251.

Oard, D. W., Levow, G.-A., & Cabezas, C. (2001). CLEF experiments at the University of Maryland: Statistical stemming and backoff translation strategies. In *Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 2000, Revised Papers*, Vol. LNCS 2069 of *Lecture Notes in Computer Science*, pp. 176–187 Heidelberg. Springer-Verlag.

Peters, C. (Ed.). (2001). *Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers.* Lecture Notes in Computer Science 2069. Springer, Berlin.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 55–63.

Ponte, J. M., & Croft, B. W. (1997). Text segmentation by topic. In *Proceedings of the 1st European Conference on Research and Advanced Technology for Digital Libraries*, pp. 113–125.

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

Resnik, P., Oard, D. W., & Levow, G.-A. (2001). Improved cross-language retrieval using backoff translation. In *Proceedings of the First Human Language Technology Conference (HLT) 2001*.

Sheridan, P., Wechsler, M., & Schäuble, P. (1997). Cross-language speech retrieval: Establishing a baseline performance. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 99–108.

Singhal, A., & Pereira, F. (1999). Document expansion for speech retrieval. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, pp. 26–33.

W.B.Frakes, & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms.* Prentice Hall, Englewood Cliffs, NJ.

Wilkinson, R. (1997). Chinese document retrieval at TREC-6. In Voorhees, E., & Harman, D. K. (Eds.), *The Sixth Text REtrieval Conference (TREC-6)*, pp. 25–30. NIST. NIST Special Publication 500-240.

Xu, J., Fraser, A., & Weischedel, R. (2002). Empirical studies in strategies for Arabic retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 269–274.