

An Introduction to Cross-Language Information Retrieval Approaches

LIS 531 - Information Retrieval - Peishan Tsai

1. Introduction

Cross-Language Information Retrieval (CLIR) addresses the situation in which a user submits a query in one language to retrieve documents in a different language. CLIR is a subset of information retrieval (IR), and shares many of the characteristics of the general IR; but CLIR is further complicated by the cross-language aspect. IR deals with the representation, storage, retrieval, and access of a monolingual document collection; CLIR has to handle the above issues and solve the problem of mapping the query that is in one language (the source language) to the document collection that is in another (the target language). An ever active research field, a vast number of papers and studies has been published on CLIR; especially since TREC, NTCIR, and CLEF developed and made available large-scale test collections. Of the four facets to an IR system: query, collection, retrieval, and feedback, most CLIR studies focus on query and collection because of its multilingual characteristic; that is also the focus of this paper.

As new techniques and methodologies are constantly being proposed, the body of CLIR research is vast and broad. This paper does not aim to be exhaustive at reviewing every study in the field, but it aims to provide an introductory overview to some approaches that deals with CLIR's cross-language facet.

Methods of CLIR always rely on some source of information (Sheridan and Bellerini, 1996). The following paragraphs will focus on some of the CLIR methodologies grouped by the resource they use to map from source language to target language. The paper will introduce machine translation approaches, dictionary-based approaches, latent semantic indexing, probabilistic-based approaches, and methods used when there is a lack of lexicon resources.

Although morphology analysis, string matching techniques, such as n-gram matching, retrieval and ranking strategy, and user interaction are also important facets of CLIR, they are beyond the scope of this paper.

2. Machine-Based Approaches

Machine Translation (MT) is the process that utilizes computer software to convert free text from one language to another; the output seeks to be accurate and fluent for human consumption. One would intuitively assume that MT is the solution to CLIR: An MT system can be used to translate the query, the document, or both into the same language, and the retrieval process could then be treated with a general IR system. The question then is which should the system translate and why.

Some argues that document translation would yield better results than query translation in that documents are generally much longer than queries, therefore they are able to provide more linguistic contexts for accurate translation (Oard, 1998).

McCarley (1999) examined the effectiveness of using an MT system by comparing the performances of three MT based systems and a monolingual IR system. The MT based systems conduct the translations in three ways: on the document collection, on the queries, and in a hybrid manner - the probability of a document being relevant to a query is computed with both

normalized probabilities of query and document translation. The same statistical translation model was used for the three models; the only difference among them is what was translated. The results showed that there is no clear advantage over either the query or document translation system, but the hybrid model surpassed both, and is comparable to the monolingual system.

Oard (1998) expanded on the study and again compared the effectiveness of query translation with document translation. Although the results showed that document translation receives higher average precision than query translation, both are still below monolingual retrieval; and the results are not clearly statistically significant because of the small query sample size.

As promising as document translation may seem, the application of the method requires full translation of the document collection; a requirement that could be computationally costly. Oard (1998) spent 10 machine-months translating a German collection that consists of nearly 252,000 newswire articles, 268 of which were unable to be translated. The time needed would be further magnified if there were multiple source languages, and when documents are frequently added to the collections; such as it is in the Web environment.

Fujii and Ishikawa (2000) proposed a two-stage method to minimize the MT document translation computational cost: the queries are translated and submitted, retrieving a set of documents in the target language; the documents are machine translated into the source language and re-ranked based on the translation. The results show that re-ranking the translated documents brought a visible improvement to the average precision, and could greatly improve the retrieval precision of an otherwise poor query translation. However, the study did not provide a baseline for comparison, so it is hard to say how the method compares to other techniques.

A few issues have so far prevented MT CLIR systems to gain popularity.

McCarley (1999) and Oard (1998) both observed that MT systems' performances vary when dealing with different languages. The translations were noticeably better when translating in a certain direction (in their cases from French to English; and from English to German) than when the translation direction is reversed (from English to French; and from German to English). The difference might be caused by different morphological analysis performed on each language and the quality of the translation model training data. This raises questions to the validity of the studies: if the results could be repeated on different language pairs or when the training data is significantly dissimilar to the collection.

MT systems require time and resources to develop; they are still not widely or readily available for many language pairs. Ballesteros and Croft (1997) pointed out past study results that indicate improvements gained by MT techniques may not outweigh the cost.

Oard(1998) and Fujii and Ishikawa (2000) also compared MT systems with bilingual lexicon translations, and found lexicon translations are comparable, if not better, than machine translations. The result could be because queries in general lack the context information MT needs to provide the accurate translation, whereas lexicons are able to generate a list of potential translations, therefore have a higher possibility of offering the right translation. On this note, the next section will look at the dictionary-based approach in more details.

2. Dictionary-Based Approaches

Dictionary-based approaches utilize machine readable dictionaries (MRD), bilingual word lists, or other lexicon resources to translate the query terms by replacing them with their target language equivalents.

Hull and Grefenstette (1996) compared the effectiveness of a monolingual IR system, and CLIR systems that translate the queries using either an automatically constructed bilingual MRD, or a manually constructed dictionary. The results show that using only MRD, can lead to a drop of 40-60% in effectiveness below that of monolingual retrieval. But a manually constructed multi-word phrase dictionary can perform as well as monolingual system. Although the study is rather preliminary, with the dictionaries being manually revised and structured, it still shows that with the correct translations of multi-word expressions, a CLIR system could perform just as well as a monolingual system. The study shows that the recognition and translation of multi-word expressions, and phrases, are crucial to success in CLIR.

Another observation is that lexicon ambiguity is a great cause for translation errors. A word can often be translated into several meanings; not all were intended in the query. A CLIR system will have either use all the translations or be able to choose among the options and find the ones that best represent the original query. One way to deal with this issue by assuming the first definition listed in the dictionary is the most frequently used, therefore selecting the terms corresponding to the first sense, or just the first term as translation (Oard, 1998). Bellesteros and Croft (1998) tested the first sense method, and found that it only brought an insignificant improvement to the average precision. Oard (1998), on the other hand, showed that selecting a random translation from multiple translations can be as effective as retaining every possible translation for a query, although both are far below the performance of monolingual retrieval.

Extraneous definitions add noise to the retrieval process because they could unbalance the query term values by giving more weight to the one with multiple translations, and devalue the query term with few or single translations. This is seen in Hull and Grefenstette (1996), and demonstrated in Ballesteros and Croft (1998), in which the noisy translations may have caused the retrieval precision to be 60% below that of monolingual retrieval. Ballesteros and Croft (1998) amended the situation by using structuring the queries with a synonym operator¹. The operator wraps the multiple translations of one query term into one unit, and treats it as a pseudo-term with only one belief value assigned for the whole package. The method brought a 47% improvement to the original result.

In addition to phrase identification and translation and the inherently ambiguity of language translation, Pirkola et al (2001) and Bellesteros and Croft (1997) point out other problems of dictionary-based translations, including: untranslatable words, such as proper names, compound words, and domain special terms that are not included in the dictionary that was used; and inflected words, which could usually be handled by stemming. Xu and Weishedel (2000) suggested that missing lexicons poses the biggest threat to CLIR system performances. That might be, but many efforts have been put into solving translation ambiguity. The next section will introduce some of the disambiguation solutions, and section 4 will introduce efforts made to

¹ The operator mentioned is the #syn operator in INQUERY's query language. INQUERY is an information retrieval system based on a probabilistic retrieval model called the inference net. For detail description, please see Broglio, Callan, and Croft (1994). For how query structuring is used in CLIR, please see Pirkola, 1998.

broaden lexicon resources. As to morphological processing and string matching techniques, they are out of the scope of this paper, and will be discussed in the future.

2.1 Disambiguation Techniques

When the query terms can be translated into different meanings in the target language, the various translations can introduce noise to the retrieval process, and harm the precision of the results. Bear in mind, though, the value of the noise is still in debate; the extra terms might actually increase the recall of a query. For example, Hiemstra and de Jong (1999) found that, sometimes, using all translation possibilities yield better results; and the quality of the retrieval relies on good search methods, not on disambiguation. Yet many other researchers found disambiguation to be valuable to the retrieval process, and many disambiguation techniques have been developed for dictionary-based methods to improve its retrieval effectiveness. Among them are part-of-speech tagging, parallel-corpus based techniques, and query expansion techniques.

2.1.1 Part-of-Speech Tagging

The concept of part-of-speech tagging for term disambiguation is to use the part-of-speech tags as the pre-selection criteria to weed out the translations that are less likely to be the equivalent of the query. The query terms are tagged with part-of-speech; among their possible translations, only the ones that have matching part-of-speech tags are chosen for further consideration. Part-of-speech tagging is often used as an initial step in other disambiguation methods, such as parallel corpus techniques (Ballesteros and Croft, 1998; Davis, 1996; Davis and Ogden, 1997; Lin, Jin, and Chia, 2005).

2.1.2 Parallel Corpora

Parallel corpora are sets of translation-equivalent texts; the corpus in language A mirrors the content and the structure of the corpus in language B. Parallel corpora are often used to determine the relationships, such as co-occurrences, between terms of different languages, and can be employed to train a statistic translation model (Chen, Bian and Lin, 1999; Gao et al, 2001).

In Ballesteros and Croft (1998), co-occurrence statistic is used for disambiguation based on the concept that correct translations of query terms should co-occur in the text and incorrect translations should not. The translations are first filtered with part-of-speech tags. Each translation candidate of a query term is then paired up with a translation candidate for another query term. Each pair's pattern of co-occurrence is calculated, and the ones with the highest co-occurrence values are chosen as the query translation. This method can be used to disambiguate single words as well as multi-word phrases.

Davis (1996, 1998) and Davis and Ogden (1997) used parallel corpus for linear disambiguation of term equivalents. After weeding out some of the translation options with part-of-speech tagging, the query terms and their translation equivalents each retrieves a set of documents from their individual language side of the parallel corpus. The translations whose retrieved sets most match those of the query terms are chosen as the correct translation. Davis (1996) reported an average retrieval precision 73.5% of monolingual IR systems'.

Ballesteros and Croft (1998) also used parallel corpus to evaluate the belief value of the translation candidates. They modified the method by Davis and Ogden (1997); instead of

comparing multiple retrieval sets for best matched translations, only one document set is retrieved from the corpus to find the best translations for each query terms.

The query terms are tagged with part-of-speech and translated into the target language. Before translation, the terms are also used to retrieve a document set from the source language side of the parallel corpus. With the retrieved set are the corresponding documents in the target language; a list of 5000 terms is extracted from them. The translation candidates of the query terms are looked up from the 5000 terms, and ranked by their positions in the list. The ones ranked highest are chosen as the query translation. If none of the candidates were included in the 5000 terms, then all of them were seen as query translations. The method only improves the average precision rate moderately. Its limited effectiveness is speculated to be caused by the parallel corpus' narrow scope.

Nie et al. (1999) tested their probabilistic translation model with parallel corpora, and observed that the translations reflect the peculiarities of the training corpus, which sometimes lead to odd translations. Furthermore, some words are not included in the training corpus; or if the word was present, its frequency in the corpus did not represent its general usage. They also saw that at times the probabilistic model fails to choose the correct translations because of the noise induced by statistical association: unimportant options could be deemed highly relevant because of a higher occurrence rate. Rogati and Yang (2004) also examined the affect of parallel corpora selection with a probabilistic translation model, and found that a mismatch of domain between the corpora and the target collection has a negative impact on the retrieval performance.

Maeda et al (2000) saw positive effect in using a domain matching parallel corpus. They tested a parallel corpus that is identical to the target collection and achieved 99% of the monolingual retrieval precision average. Yet being aware of how impractical it is to prepare a comprehensive corpus to cover all possible domains, they proposed using the World Wide Web as a multilingual corpus. The Web based method is described in later paragraphs.

McNamee and Mayfield (2002) confirmed and quantified the degree to which inferior lexicon resources affect dictionary-based and corpora-based techniques by intentionally degrading the parallel corpora and bilingual wordlists prior to use. Kraaij (2001) offered the opinion that the mean average retrieval precision is proportional to the lexical coverage of the corpora or dictionary.

In addition to the problem of domain discrepancy between training corpus and document collections, there are other major drawbacks to using parallel corpus: they are hard to come by, and difficult to develop. Those that are available may be small in size or are narrow in subjects. The texts do not reflect the document collection, the query term meanings, or general lexicon usage (Ballesteros and Croft 1998, 1997; Pirkola et al. 2001; Gao et al. 2001).

2.1.3 Query Expansion

Local feedback and local context analysis are two popular query expansion methods used by information retrieval systems to solve word sense mismatch problem. The problem arises when the queries and the documents contained different words to describe the same concept. The discrepancy of words could cause a relevant document to be missed at retrieval.

Local feedback is also known as pseudo relevance feedback or blind feedback. It assumes the initial top-retrieved documents are relevant, and instead of returning the documents back to the users, extracts additional relevant terms from them to expand the query (Xu and Croft, 2000).

Local context analysis is a method proposed by Xu and Croft (2000) that employs co-occurrence analysis for query expansion. Concepts, instead of terms, are extracted from the top-retrieved documents. The concepts could be made up of single words or multi-word phrases. They are ranked according to their co-occurrence rate with the query terms in the retrieval set, and the higher ranked concepts are used for query expansion. It has been shown to be more effective than local feedback.

Ballesteros and Croft (1997) explored the efficacy of reducing dictionary-based translation errors with the two query expansion methods. They applied the expansions before, after, or both before and after query translation.

Pre-translation feedback is performed with the original query in a source language database. The assumption is that the expansion terms are able to provide extra context as anchors for disambiguation; therefore creating a stronger base for translation, and improving the overall precision. However, the extra context could also introduce inappropriate translation term to harm the precision.

Post-translation feedback is performed with the translated and disambiguated query terms in a target language database. The assumption is that the additional terms added from the expansion can de-emphasize irrelevant translations therefore reduce ambiguity and improving precision; it can also improve recall by broadening the query with other relative terms.

Ballesteros and Croft (1997) found that query expansion via local feedback and local context analysis are able to significantly reduce the number of dictionary translation errors. In general, local context analysis, which expands query terms with multi-term phrases, generates a higher precision than local feedback either at pre- or post-translation, although the recall levels are lower. The best result with both local feedback and local content analysis were achieved when both pre-and post-translation feedback are used. The study found that the errors of automatic translation were reduced by 45% when local content analysis was used pre- and post-translation.

Ballesteros and Croft (1998) focused on local context analysis, and examined: the effects of combining pre-translation feedback with co-occurrence-statistics disambiguation techniques, word by word translation versus phrasal translation, as well as the effect of combining post-translation feedback with query structuring. Co-occurrence statistic is described in more detail in section 2.1.2. Query structuring helps to contain translation ambiguity by normalizing the variation in the number of translation equivalents across query terms. When the query terms have multiple translations, and all potential translations included in the new query, each with a unique weight assignment, the original query terms with more translations receives more weight as a result, consequently distorting the original query sense. But when the translations are structured as synonyms, they are seen as one pseudo term and given one weight, thus eliminating the effect of biased weight assignment.

Ballesteros and Croft (1998) shows that the impact of pre-translation analysis lessens as query disambiguation improves, but query expansion may still be useful in providing anchors for disambiguation through co-occurrence technique. Post-translation expansion, either used alone or with pre-translation expansion, can enhance both recall and precision. The best retrieval result was achieved when various disambiguation methods were integrated. Combining either of the query expansion methods with phrasal translation and co-occurrence disambiguation methods can achieve 90% of the monolingual retrieval results.

McNamee and Mayfield (2002, 2004) further demonstrated that pre-translation expansion is able to achieve good performance when only very poor linguistic resources are available, indicating that low density language will greatly benefit from pre-translation expansion method.

3. Latent Semantic Indexing (LSI)

LSI (Littman, Dumais and Landauer, 1998) is a variant of the vector-space model. The central idea is that term-term inter-relationships can be automatically modeled and reflected in a vector-space. LSI uses a linear algebra, singular value decomposition, to discover the associative relationships between terms.

LSI does not rely on external lexicon resources, such as MRD, to determine word relationships; the relationships are derived from a numerical analysis of the initial training data instead. The training data is usually a set of multilingual documents. LSI method ignores the word orders and treats the documents as a bag of words. The method examines the similarity of the contexts in which words appear, and creates a reduced-dimension feature-space representation. In this lexicon dimension, words used in similar contexts are located close together. Documents are also represented in the same vector space, therefore similarities between any combinations of words and documents can be obtained.

LSI is unique in many ways. First of all, all terms are treated as related instead of independent, as are in most other methods. Because the terms are relative, LSI is able to retrieve relevant documents even when the documents and the queries do not share the same terms. Once the model is established, new materials could be added in at any point without model re-establishment or adjustment. New documents can be folded into the model as long as the existing dimension space is a reasonable characterization of the new items, and that the items can be represented in it. The method is entirely algorithmic, and does not need other resources besides the initial training data for retrieval or translation.

Evans, et al (1998) demonstrated the power of LSI by using the method to map variants of medical concept expressions and terminologies. They noticed several important features to the method: it does not depend on explicit semantic representations or on word-for-word corresponding among words; the initial training data can be quickly developed; the model is also tolerant of noise and fuzzy approximations of concepts.

Nevertheless, there are some drawbacks to this method as well (Evans et al, 1998). The learned associations between terms are specific to the domain of interest. Words that are used with different senses will result in semantic distortions. LSI is also computationally expensive, and may be quite costly when dealing with a larger data set. These problems have prevented the wider application of LSI.

Chau and Yeh (2002) propose a method that is similar in concept to LSI: the fuzzy keyword classification, in which terms and documents are represented in a concept space for similarity evaluation.

In fuzzy keyword classification, keywords are extracted from a comparable or parallel corpus to establish a multilingual concept directory. Fuzzy clustering algorithm is then applied to group conceptually related keywords into concept classes. It is a fuzzy algorithm because each keyword can be classified to more than one class, and is assigned a membership value to the classes. With the fuzzy multilingual keyword classification scheme as the concept directory, the documents are then each mapped to the concept classes that they belong. A similarity measure

between the document and the concept space is computed to organize the documents. The study does not provide quantitative evidence to the effectiveness of the method, but the clustering is able to provide a contextual overview of the document collection for exploratory searching and document browsing. And, as does LSI, it provides a solution to the vocabulary mismatch problem between query concept and documents.

3. Probabilistic-Based Approaches

Probabilistic-based approaches are statistic systems that use algorithms to predict query matching, document relevance, and document belief value. The approaches include: corpus-based methods, which translates queries, and language modeling, which seeks to bypass translation.

3.1 Corpus-Based Approaches

There are two types of multilingual corpora: parallel corpora that have translation-equivalent texts; and comparable corpora, in which texts of the same subject are not aligned, nor direct translations of each other, but composed in their respective languages independently.

3.1.1 Parallel Corpora

Parallel corpora's translation-equivalent characteristic allows for the mapping of equal terms between languages. When the texts are aligned, parallel corpora provide the resources to determine the correlation of words in different languages, and the probability of one term being the translation equivalent of another. As section 2.1.2 indicated, parallel corpora are often used for term disambiguation and query expansions. They are also often used as training material for statistical machine translation systems.

One example is the "fast document translation" system by IBM (Franz, McCarley, and Roukos, 1999). The system built bilingual dictionaries and translation models using algorithms automatically learned from aligned texts of parallel corpora. With the system, the translation can be done "within an order of magnitude of the indexing time" (Franz, McCarley, and Roukos, 1999, 157). The system was incorporated with a general IR system, and was shown to be highly effective in CLIR testing.

Another example of parallel-corpora based system is HAIRCUT (McNamee and Mayfield, 2004). The system uses parallel corpora to develop a statistic model for n-gram based retrieval technique. The technique relies on language similarity instead of direct translation for query term mapping, and was shown to be highly effective in CLIR testing.

McNamee and Mayfield (2004) described their method as language-neutral methods because it is not limited to one translation direction: it can translate language A to language B, and from language B to language A without modification. However, the method can only be used among languages with similar structures, such as among European languages, or among certain Asian languages.

The advantage of these methods is that neither of them are language dependent. The models can be adapted to any language as long as there are sufficient training materials provided for the language. But as Franz, McCarley, and Roukos (1999) pointed out, with linguistic resources varying widely both in size and quality for different languages, it is necessary to develop separate systems for each language pair in order to factor in the training data variables.

The aforementioned studies did not address the issue of corpora domain, which has seen to influence the outcomes of corpora-based approaches at word sense disambiguation. Would corpora domain be one of the training data variables to affect translation results? Further exploration is needed to answer that question.

3.1.2 Comparable Corpora

Because the corpora contents are written in their individual languages for their respective readers, comparable corpora provide a data source for natural language lexical equivalents.

Sheridan and Ballerini (1996) used comparable corpora to generate a similarity thesaurus for CLIR. A similarity thesaurus is constructed by extracting terms from documents, and grouping them together based on the concept they represent in the texts; it is used for query expansion in general IR. For CLIR, the similarity thesaurus is multilingual. When a query is submitted, it is expanded with the use of the thesaurus to contain similar terms in all languages. From the expanded query, terms in the target language are filtered and submitted to the database for document retrieval.

The query expansion process is not only able to provide query translation equivalents, but also able to increase recall by adding terms of similar concepts to the query. On the other hand, the expanded query may introduce terms that have a lower degree of relation, thus add noise to the query and hurt the retrieval precision.

Picchi and Peters (1998) extended a comparable corpus processing procedure to CLIR. The system was originally designed to retrieve comparable texts in different languages. The basic idea behind the method is not to extract the precise translation equivalent, but to find the set of texts that has the highest probability of corresponding to the texts written in another language. How are the similar contexts identified? When two texts are similar, it is likely that several of their components are also equivalents. Once the equivalents are found, they become the breadcrumbs that would lead to similar texts. With this in mind, Picchi and Peters (1998) uses a lexical database of comparable corpora, accompanied with morphological procedures to obtain co-occurrence and correlations among terms, for query translation.

Franz, McCarley, and Roukos (1999) used comparable corpora to train a probabilistic based CLIR system when parallel corpora are not available. The study aligned the comparable passages and treated them as parallel corpora. From the aligned texts, the system was able to extract bilingual word-pairs and use them for CLIR tasks.

As are with parallel-corpus based methods, the aforementioned methods are also language independent. Similarity thesaurus can be expanded to include multiple languages with additional comparable corpora texts; it is also multi-directional; the languages to be translated and for translation are interchangeable. Once the similarity thesaurus is constructed, it can take queries from any of the languages it covers, and return a retrieval set in the rest of the languages.

The drawback of the approaches is its reliance on corpora. Though it is presumed that comparable corpora are easier to obtain than parallel corpora, it may still be hard to develop or to acquire a large enough set. It is also domain specific. As is LSI, a term with different usages may skew the methods or disrupt the thesaurus and hinder the retrieval effectiveness.

Gao et al (2002) proposed an alternative model that can be trained with unrelated corpora. The triple translations model extends the basic co-occurrence model by incorporating in syntactic

relations between words. When words are used in a text, there is a syntactic relation between every adjacent pair that can be described as: (word1, syntactic relationship, word2), called the triple. These strong syntactic dependencies in the original language usually remain after translation. For example, in the phrase “big fish”, the words big and fish have an adjective-noun relationship; as a result, the translation of the phrase will most likely be made up of two terms with an adjective-noun relationship. Therefore, among all translation candidates for the query terms, the best combinations will be the ones that have the highest likelihood of being used in the same syntactic manner, forming a similar triple.

The advantage of the triple translation model is that it does not rely on parallel nor comparable corpora. The model only requires the estimation of the triple probabilities for each language, and it could be done separately, therefore, it could be trained with a set of unrelated corpora. However, Gao et al (2002) only tested one language pair, English and Chinese, with the model. As disparate as the two languages are, they share a somewhat similar syntactic structure. It would be interesting to find out how language dissimilarities would affect the outcome.

3.2 Language Models

Language modeling is used in information retrieval to predict the occurrences of terms in a document without regard to sequential orders (Ponte and Croft, 1998). For CLIR, it is used to model the generation of a query in one language, given the document in another.

As Larkey and Connell (2004, 460) described: “A query is a bag or a sequence of single terms, generated from independent random samples of a term from one of two distributions – the distribution of words in a model of a document, and the distribution of words in a background model such as General English.” Where traditional probabilistic models estimate the possibility of a document being relevant given a query, language models assume that users have a general idea of what terms are likely to be found in their target documents. So given a query, the document model, and the background language model, language models can estimate the probability of the query being generated from any of the documents in the collection. Several different cross-language language models have been proposed (Lakey and Connell, 2005); the following paragraphs will introduce some of them.

Twenty-One is an information retrieval project that ran from 1996 to 1999; it was funded by European Union Telematics Applications programme, sector Information Engineering (Hiemstra et al., 2001). All of its information retrieval tasks, whether monolingual or cross-language, were carried out based on a single unigram language model. In the model, the probability of single and stemmed query terms are used to generate a document-relevance score (Hiemstra et al. 2001; Hiemstra and Kraaij, 1999). In this model, the query formulation process and the translation process are integrated: the queries are translated using a probabilistic dictionary. In a probabilistic dictionary, term translations, s , and their probability of occurrence in the document selection, t , are listed together as pairs, (s, t) . When there is more than one translation that has a possibility of occurrence, the possible translations are grouped together, forming a structured query. But because the document collection is not translated, the translation possibilities will have to be estimated from other resources, such as from a parallel corpus. And when a query term translation is not seen in the lexicon resource, the model is interpolated with a background language model to compensate for the data sparseness.

Hiemstra et al. (2001) also proposed a new relevance feedback method. It follows the spirit of traditional relevance feedback, but instead of using the initial retrieved documents for query expansion, the documents are used to re-estimate the translation probabilities and the importance of each query term. The re-estimating model did not result in improvement of retrieval performance in the study; however, it showed great potential at processing user-feedback in an interactive CLIR setting.

Xu and Weischedel (2000) and Xu, Weischedel, and Nguyen (2001) propose using Hidden Markov Model to simulate the query generation process. The documents that have a higher probability of generating the queries are deemed relevant.

The studies used a bilingual lexicon as well as parallel corpus to estimate the translation probabilities. When only a bilingual dictionary is available, it is used to obtain the terms and their corresponding translations; in which case, the same translation probabilities are assigned to all translation candidates of a word. When parallel corpus is available, the model uses the texts to estimate translation probabilities. The study found that the best results were achieved when the two lexicon resources are used together. Xu, Weischedel, and Nguyen (2001) concluded that the performance of Hidden Markov Model is comparative to other CLIR methods, and slightly more effective than MT systems. However, they observed that how the parallel corpora match up with the document collection strongly influences the retrieval outcome. The study used corpora with texts in a dialect when the document collection was not, and the variation between languages brought negative impact to the retrieval results.

Lavrenko, Choquette and Croft (2002) suggested a relevance model that bypasses translation altogether. The model estimates the joint probabilities of each individual word in the target language co-occurring with the query terms. The model takes into consideration the relevance of the entire target language vocabulary, a form of query expansion; as well as each word's co-occurrence with the query terms, a form of term disambiguation. One can say that the model has built in query expansion and term disambiguation mechanism. The model is different from other language models in that the previous approaches make use of translation probabilities attached to pairs of words, whereas the relevance model does not rely on word-by-word translation when parallel or comparable corpora are available. The model first retrieves one set of matching documents in the source language from the corpora, it then uses the set of comparable documents in the target language to estimate, for every word in the target language vocabulary, the probability of observing the word in the set of relevant documents. Retrieved documents are then ranked by divergence, where the documents with less divergence from the query terms are deemed more relevant.

Language model approaches have become popular in recent years because it has a firm foundation of statistical theory, is language independent, and can easily incorporate in additional enhancements, such as document expansion, stemming alternatives, etc (Larkey and Connell, 2005). On the other hand, there are still several issues that require further exploration.

Xu and Weischedel (2005) measured CLIR system performances with different lexicon resources, and concluded that while using bilingual term list achieves acceptable retrieval results, combining the term list with parallel corpora produces a result comparable to that of monolingual systems. They also observed that pseudo-parallel text produced by machine translation can partially substitute parallel text. Lavrenko, Choquette and Croft (2002) pointed out that lexicon coverage is extremely important for accurate translation probability estimations; the model still

relies on good parallel corpus for effective retrieval results. Hiemstra et al. (2001) countered the problem by mining Web documents to develop parallel corpora for their study; but the vocabulary coverage did not match up with the query terms, resulting in incorrect probability estimations.

Another interesting issue lies in the key component of language models: the estimation of translation probabilities. It is assumed that good estimations of the probabilities are the determining factor in the effectiveness of language modeling; but researchers have observed that large changes to translation probabilities made little difference to the retrieval effectiveness (Larkey and Connell, 2004).

4. When there is a Lack of Lexicon Resources

However sophisticated the aforementioned CLIR approaches are, they all share one weakness: the reliance on some kind of lexicon resources, such as machine readable bilingual dictionaries or parallel corpora, for probability assessment or translation. But not all languages have such resources readily available. Researchers have come up with different alternative methods when faced with this limitation.

4.1 Transitive and Triangulation Methods

Transitive and triangulation methods use other languages that have sufficient lexicon resources to accommodate for the lack of bilingual translation resources between a language pair. These methods not only allowed for CLIR between languages that do not share translation resources; but also reduce the number of individual translations needed when translations have to be performed between a large number of languages.

4.1.1 Transitive Methods

Transitive methods use a medium language to bridge the translation gap between two languages.

Franz, McCarley, and Roukos (1999) merged two translation system (from language A to language B, and from language B to language C) to obtain a new translation system (from language A to language C). Another model they introduced is a mergence of two CLIR systems. Queries are submitted in language A; the query retrieves a document set in language B; a new query is formed in language B with the retrieved set, and used in a CLIR system between language B and C.

Kishida and Kando (2005) examined a hybrid approach that translates both queries and documents to an intermediary pivot language; in this case, English. The queries are translated by a MT system, and the documents are “pseudo-translated” by replacing source language texts with term translations directly plucked from a bilingual dictionary. The retrieval process begins with translating the query into the intermediary language then the target language; with the translations, the system retrieves a set of documents in the target language. At the same time, the documents are roughly translated into the intermediary language, and a set of the translated documents are retrieved with the query in the intermediary language. The two sets of retrieved documents are merged for the final result. Unfortunately, the hybrid approach could not outperform other approaches. The reason for the low performance may be the lower quality of the machine readable dictionary used, and the omission of translation disambiguation in the process.

Lehtokangas, Airio, and Jarvelin (2004) evaluated the performance of transitive translation using a simple dictionary translation in combination with other techniques, including: morphological analyzers, stop-word list, query structuring, n-gramming matching for un-translatable words, and triangulation (to be described below). The results showed that additional techniques are able to enhance the retrieval results except for triangulation, which becomes unnecessary or even harmful when query structuring is used. The study findings were encouraging; it showed that transitive translation incorporated with query structuring was able to perform comparably, if not better, to that of traditional bilingual retrieval. Even so, there are still several research questions to be answered.

Lehtokangas, Airio, and Jarvelin (2004) used machine readable dictionaries from the same publisher for the translation task, which may have a favorable effect on the transitive effort. Would different lexicon resources reproduce the same positive outcomes? The experiments are conducted with English, a language rich in lexicon resources, as the target language; and European languages that are from the same language family as source and intermediate languages. The similarity between languages and the resources available would have facilitated translations. How would different language pairs or dissimilarities between languages affect the retrieval result? If the transitive approach with one intermediate language performed close to bilingual CLIR results, what would happen if additional intermediate languages were added? Would extra translation procedures introduce error and more ambiguity into the process?

4.1.2 Triangulation Methods

Gollins and Sanderson (2001) presented a CLIR approach using lexical triangulation, in which translations from two different transitive routes are used to extract the translations in a third language. Assume a user wants to retrieve a document set in language Y with a query formed in language X, but there are no direct translation resources available between X and Y, so two intermediary languages, A and B, are used to aid the process. The query terms in X are translated into both A and B. The translations in A and B are then all translated into Y, yielding two sets of translations. The sets of translations are matched up and compared. Only the union of the two sets, which is terms seen in both translation sets, is kept as the final queries. In essence, triangulation method generates two transitive translations, and cancels out the extra noise by comparing the two different results of the same translation.

Gollins and Sanderson (2001) first compared the performance of bilingual CLIR, triangulation, transitive, and monolingual IR systems. Without additional techniques, such as query expansion, none of the approaches are able to compare with monolingual IR. Among them, direct bilingual CLIR outshines the rest, and triangulation method performs slightly better than transitive approach. However, when extra intermediary languages were added to the triangulation effort, adding additional comparison sets for query forming, the results showed an increase in recall. The improvement might have led to an improvement in the retrieval results, making it good enough to rival those of direct bilingual translations.

Gollins and Sanders (2001), however, noticed that different language pairings appear to affect the outcomes of triangulation translation. But the effects they observed could be more from the relative size of the lexicon resources, than from the property of the languages themselves.

Ballesteros and Sanderson (2003) compared transitive and triangulation methods, and examined the effectiveness of query disambiguation by employing query structuring with triangulation

methods. Triangulation approach is able to eliminate the ambiguity introduced by using pivot languages, and query structuring eliminates ambiguity in the original query terms. The study tested a number of test collections and language pairs, and showed that the combination of query structuring and triangulation translation is an effective CLIR methodology, and performs better than both transitive translation and direct bilingual retrieval.

As with all other methods, there are weaknesses to triangulation method as well. One is encountered when there are no common words in the translation sets. The methods either uses all options in the translation sets (the liberal approach in Gollins and Sanders, 2001), or pass along the original query term (the strict approach in Gollins and Sanders, 2001; Ballesteros and Sanderson, 2003). The first approach creates too much ambiguity and noise to the translation, and the latter may cause a zero fit between the query terms and the document contents. Another problem is that non-intersecting query words are dropped from the translation, which may lead to losing some of the query concepts.

4.2 With the Web as Resource

As previously mentioned, while many of the CLIR approaches rely on parallel or comparable corpora for system training or translation, these resources are hard to come by. Many researchers have since turned their eyes to the Web, where an abundance of bilingual or multilingual websites lie in wait to be used for parallel text construction. Such efforts are seen in the STRAND system developed by Resnik and Smith (2003), and PTMiner developed by Nie and Cai (2001). These approaches have been put to test in CLIR tasks.

Nie et al. (1999) tested a probabilistic translation model with training corpus automatically gathered from the Web. The corpus constructing technique is similar to that of PTMiner that was developed later. The Web was crawled and harvested; candidate parallel webpage pairs are identified by similar URL structure, HTML structure of the webpage, and text alignments. The corpus is then used to train a probabilistic model, and used for CLIR. The researchers observed some particular problems with Web constructed parallel corpus. For example, geographic names are often seen to be listed as the translation of another geographic name, because it is common for websites to contain lists of locations, resulting in high co-occurrence rates for the geographic names, and be deemed relevant by the statistic system. Another problem is that terms in the source language often appear as the target language translation, because some of the terms were left un-translated on the supposedly parallel webpage. The quality of Website translations varies, and the webpage contents are often noisy. Nie et al. (1999) found that combining a bilingual dictionary with the Web constructed parallel corpora greatly improves the retrieval results, and reaches above 80% of monolingual IR results.

Chen and Nie (2000) and Kraai, Nie, and Simard (2003) tested statistical models with PTMiner. PTMiner utilized existing search engines to search for candidate sites for corpus building. The candidate sites are those that provide anchor text links to other same content multilingual webpages. The webpages are then paired up as translation candidates by URL comparison, text length, HTML structure and alignment, etc. Pages that were found parallel are downloaded to construct the corpus. Both studies found that the parallel corpora constructed have a higher degree of noisiness, where the translations are not exact, but still related to the original query. The noises can create a query expansion effect, and increase the retrieval recall. Chen and Nie (2000) also found that combining bilingual dictionary with the Web corpora can greatly improve the retrieval precision; however, the system was only able to achieve near 70% of the

monolingual results. The low performances could be because of the divergence between languages that causes difficulties in text alignment. In addition, as Nie et al. (1999) noticed, the noise that created a query expansion effect, and the varying translations quality of Websites could also hinder the translation effort by adding incorrect terms into the query. Finally, as all parallel corpora-based approaches have showed, different domain coverage of the training material and the document collection have great negative impact on the translation result, where terms are either translated incorrectly, or untranslatable all together.

Lu, Chien, and Lee (2004) proposed constructing multilingual lexicons by mining Web anchor texts and link structures. Anchor text is the text on a webpage that links to another webpage; it is usually a brief description of the webpage that it links to, including titles, headings, etc. When anchor texts in different languages on different webpages link to the same destination, chances are, they are expressing the same concepts. Therefore, a collection of anchor-text set can be considered as a comparable corpus. But because not every combination of language pairs has sufficient anchor texts for query translation, Lu, Chien, and Lee (2004) proposed a transitive translation model for general application, where a language more commonly used for Web contents is chosen as the pivot language. Their study showed that using anchor text alone is indeed not enough; the small corpus size could greatly harm retrieval results. With a more popular language as intermediary, the chances of finding corresponding anchor texts increase, consequently improving the translation rate and achieve higher precision for the retrieval. Despite the improvement with the transitive approach, it is still questionable whether there is enough anchor text to support all query needs. Would texts used in different context and domain affect the query translation? How about the different qualities of website content?

Instead of building parallel corpora, Zhang and Vines (2004) and Chen et al. (2004) used search engines to dynamically generate parallel text after a query is submitted. When a query is submitted, the CLIR systems send it to online search engines, such as Google, to retrieve relevant webpages in the target language. Often times, the key terms will be in the summary description of the search results, so the summary texts are extracted and processed through a probabilistic model to find query translations. They found that the Web's timely reflection to new technology, world events, and popular culture makes it an ideal resource for translating proper names, new technical terms, and other out of vocabulary terms.

Maeda et al (2000) treated the Web as a multilingual corpus, and used search engines to obtain the coefficient between query terms and translation combinations. Query terms are translated with a bilingual dictionary, and the translation candidates are paired up, one query term's translation candidate to another query term's translation candidate, as was done in Ballesteros and Croft (1998). Each of the original query terms and the translation pairs are submitted to the search engine to obtain the numbers of documents that are relevant to them. The numbers are then used to calculate the coefficient rate of the translation pairs. The pairs with a degree of coefficient exceeding a certain threshold value are selected as the target language query. The study noted, however, that the quality of the Web search results is not always consistent, and the translations search engines provide are not always suitable. It is therefore necessary to include multiple candidate pairs for caution, and also attain query expansion effect.

Kimura, Meda, and Uemura (2004) proposed employing Web directory that has versions in different languages, such as Yahoo!, as a translation resource. The categories in different languages are matched up; and feature terms are extracted to represent the categories. When a query term is submitted, the system first determines the relevant category of the query in the

source language. The query is then translated into the target language with a dictionary and disambiguated using the feature term set of the relevant category. The results were not satisfactory. The bilingual dictionary was unable to produce good translations for the query terms. And without good translations to work on, the disambiguation effort was futile. Another problem comes from insufficient term extraction from the Web; it further lowers the retrieval precision.

Larson, Gey, and Chen (2002) propose using a different online resource, library online public access catalogs (OPAC), as translation resource. Library collections are usually multilingual, with each item cataloged in the language the item is created in. The subject headings can “form interlingual mappings between pairs of languages or from one language to several languages” (Larson, Gey and Chen, 2002, 185). Furthermore, international library catalogs can be harvested with the search and retrieval protocol Z39.50; allowing the expansion of lexicon resources and inclusion of different languages. However, the system is still in development, and is now only limited to dealing with single word translations. One other issue that needs dealing with is that most OPACs in the U.S. Romanized other languages in their systems, and when foreign languages are retrieved, they are also transliterated. Once the retrieval process is complete, the characters need to be transliterated back to their original form for user consumption.

5. Conclusion

As stated in the Introduction, this paper does not seek to provide a comprehensive coverage of every CLIR methodologies and techniques available; it seeks to provide an introductory overview of the field. This paper outlined various CLIR methods: dictionary approach and the disambiguation techniques; probability/statistic-based methods; transitive and triangulation methods; and Web-based approaches.

All approaches have their strengths, and all have their weaknesses. Some are costly, such as LSI, some are rather complex with different approaches integrated within. Some seems straightforward, such as dictionary translation, some with questionable effectiveness, such as anchor text approach. All are fascinating, but few are available to the general public still.

There is no best solution for CLIR thus far. In fact, with most studies done in a controlled environment with test collections, it is unclear whether the results could be re-produced in real life search situation. Most of the studies are done with languages spoken in highly developed regions, including North America, Europe, and East Asia, where there are also the most lexicon resources, and Web development. But CLIR should not, and could not, be limited to these languages. There should be ways to access less spoken languages as well. Pirkola et al. (2002) looked into CLIR in Zulu, and found that the lack of lexicon resource made most of the existing CLIR technologies not feasible for Zulu-English retrieval. These are still issues that need to be solved.

The use of the Web seems promising as well. However, with no quality control of Web contents, one really needs to approach the resources with care.

This paper does not cover the techniques and issues of CLIR system interface and user interaction. How could a user benefit from CLIR and make use of the resources is an interesting topic for another day.

Despite the issues that are yet to be solved, the outlooks of CLIR are positive. And hopefully, it would soon be time for CLIR to be a general application as wide spread as search engines.

6. Reference

- Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swann, R., and Xu, J. (1997). INQUERY does battle with TREC-6. In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, NIST, 169--206. Retrieved on November 22, 2005, from <http://citeseer.ist.psu.edu/broglio94inquery.html>.
- 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* (Philadelphia, Pennsylvania, United States, July 27 - 31, 1997). N. J. Belkin, A. D. Narasimhalu, P. Willett, and W. Hersh, Eds. SIGIR '97. ACM Press, New York, NY, 84-91. Retrieved: 10/4/05, from ACM Portal.
- Ballesteros L., Croft, W.B., (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, 64-71. Retrieved: 10/4/05, from ACM Portal.
- Ballesteros, L. and Sanderson, M. (2003). Addressing the lack of direct translation resources for cross-language retrieval. In *Proceedings of the Twelfth international Conference on information and Knowledge Management* (New Orleans, LA, USA, November 03 - 08, 2003). CIKM '03. ACM Press, New York, NY, 147-152. Retrieved: 10/4/05, from ACM Portal.
- Broglio, J., Callan, J. P., Croft, W. B. (1994). INQUERY system overview. *Project TIPSTER Text Program, Phase I*. Retrieved on November 27, 2005, from: <http://citeseer.ist.psu.edu/broglio94inquery.html>
- Chau, R. and Yeh, C., (2002). Explorative multilingual text retrieval based on fuzzy multilingual keyword classification. In *Proceedings of the 5th International Workshop Information Retrieval with Asian Languages*, 33-40. Retrieved 10/4/05, from ACM Portal.
- Chen, H., Bian, G., and Lin, W. (1999). Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *Proceedings of the 37th Annual Meeting of the Association For Computational Linguistics on Computational Linguistics* (College Park, Maryland, June 20 - 26, 1999). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 215-222. Retrieved: 10/4/05, from ACM Portal.
- Cheng, P., Teng, J., Chen, R., Wang, J., Lu, W., and Chien, L. (2004). Translating unknown queries with web corpora for cross-language information retrieval. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 146-153. Retrieved: 10/17/05, from ACM Portal.
- Davis, M. (1996). New experiments in cross-language text retrieval at NMSU's Computing Research Lab. In the *Fifth Text Retrieval Conference (TREC-5)*, Gaithersburg, MD: National Institute of Standards and Technology, 1996, 447-453. Retrieved November 21, 2005, from http://www.scils.rutgers.edu/~muresan/IR/TREC/Proceedings/t5_proceedings/t5_proceedings.html.
- Davis, M. (1998). On the effective use of large parallel corpora in cross-language text retrieval. In G. Grefenstette ed. *Cross-Language Information Retrieval*. Kluwer Academic Publisher, 11-22.
- Davis, M. W. and Ogden, W. C. (1997). QUILT: implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Philadelphia, Pennsylvania, United States, July 27 - 31, 1997). N. J. Belkin, A. D. Narasimhalu, P. Willett, and W. Hersh, Eds. SIGIR '97. ACM Press, New York, NY, 92-98. Retrieved: 10/10/05, from ACM Portal.
- Evans, D. A., Handerson, S. K., Monarch, I. A., Pereiro, J., Delon, L., and Hersh, W. R., (1998). Mapping vocabularies using latent semantics. In G. Grefenstette ed. *Cross-Language Information Retrieval*. Kluwer Academic Publisher, 63-80.
- Franz, M., McCarley, J. S., Roukos, S., (1999). Ad Hoc and Multilingual information Retrieval at IBM. In *Proceedings of the Sixth Text REtrieval Conference (TREC-7)*, NIST. 157-168. Retrieved November 23, 2005, from http://trec.nist.gov/pubs/trec7/t7_proceedings.html.

- Fujii, A., Ishikawa, T. (2000). Applying machine translation to two-stage cross-language information retrieval. *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas (AMTA-2000)*, Oct. 2000, 13-24. Retrieved November 21, 2005, from <http://arxiv.org/abs/cs.CL/0011003>.
- Gao, J., Nie, J., Xun, E., Zhang, J., Zhou, M., and Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 96-104. Retrieved: 10/14/05, from ACM Portal.
- Gao, J., Zhou, M., Nie, J., He, H., and Chen, W. (2002). Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 183-190. Retrieved: 10/14/05, from ACM Portal.
- Gollins, T. and Sanderson, M. (2001). Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (New Orleans, Louisiana, United States). SIGIR '01. ACM Press, New York, NY, 90-95. Retrieved: 11/10/05, from ACM Portal.
- Hiemstra, D. and de Jong, F., (1999). Disambiguation strategies for cross-language information retrieval. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, 274-293. Retrieved November 23, 2005, from <http://citeseer.ist.psu.edu/hiemstra99disambiguation.html>.
- Hiemstra, D. and Kraaij, W. (1999). Twenty-One at TREC-7: Ad-hoc and cross-language track. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, NIST. Retrieved November 27, 2005, from <http://citeseer.ist.psu.edu/82362.html>
- Hiemstra, D., Kraaij, W., Pohlmann, R., and Westerveld, T. (2000). Twenty-One at CLEF-2000: Translation resources, merging strategies and relevance feedback. In *Working Notes for CLEF Workshop*. Retrieved November 27, from <http://clef.isti.cnr.it/DELOS/CLEF/Notes.html>.
- Hull, D. A. and Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Zurich, Switzerland, August 18 - 22, 1996). SIGIR '96. ACM Press, New York, NY, 49-57. Retrieved: 10/4/05, from ACM Portal.
- Kimura, F., Maeda, A., Uemura, S. (2004). CLIR using Web directory at NTCIR4. In *Working Notes of the Fourth NTCIR Workshop Meeting* (Tokyo, Japan, June 2-4, 2004). Retrieved November 27, 2005, from <http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/CLIR/NTCIR4WN-CLIR-KimuraF.pdf>.
- Kishida, K. and Kando, N. (2005). Hybrid approach of query and document translation with pivot for cross-language information retrieval. In *Working Notes for the CLEF 2005 Workshop* (Vienna, Austria, September 21-23, 2005). Retrieved November 27, 2005, from http://www.clef-campaign.org/2005/working_notes/CLEF2005WN-Contents1.htm.
- Larson, R. R., Gey, F., and Chen, A. 2002. Harvesting translingual vocabulary mappings for multilingual digital libraries. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (Portland, Oregon, USA, July 14 - 18, 2002). JCDL '02. ACM Press, New York, NY, 185-190.
- Lehtokangas, R., Airio, E., and Järvelin, K. (2004). Transitive dictionary translation challenges direct dictionary translation in CLIR. *Information Processing and Management: an International Journal*, vol. 40, no. 6, 973-988. Retrieved 10/17/05, from Elsevier.
- Lu, W., Chien, L., and Lee, H. (2004). Anchor text mining for translation of Web queries: A transitive translation approach. *ACM Transaction on Information Systems*, vol. 22, no. 2, 242-269. Retrieved at: 10/14/05, from ACM Portal.
- Kraaij, W. (2001). Comparing translation resources. In *Proceedings of the CLEF-2001 Workshop*. Retrieved November 22 2005, from <http://citeseer.ist.psu.edu/kraaij01tno.html>.
- Larkey, L. S. and Connell, M. E. (2005). Structured queries, language modeling, and relevance modeling in cross-language information retrieval. *Information Processing and Management: an International Journal*, vol. 41, no. 3, 457-473. Retrieved 11/4/05, from Elsevier.

- Lavrenko, V., Choquette, M., and Croft, W. B. (2002). Cross-lingual relevance models. In *Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 175-182. Retrieved: 11/4/05, from ACM Portal.
- Larkey, L.S., and Connell, M.E. (2005). Structured queries, language modeling, and relevance modeling in cross-language information retrieval. *Information Processing and Management*, vol. 41, no. 3, 457-473. Retrieved on 10/17/05, from Elsevier.
- Littman, M.L., Dumais, S.T. and Landauer, T.K. (1998). Automatic Cross-language Information Retrieval using Latent Semantic Indexing. In Grefenstette, G. ed. *Cross Language Information Retrieval*, Kluwer Academic Publishers, 51-62.
- Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. (2000). Query term disambiguation for Web cross-language information retrieval using a search engine. In *Proceedings of the Fifth international Workshop on on information Retrieval with Asian Languages* (Hong Kong, China, September 30 - October 01, 2000). IRAL '00. ACM Press, New York, NY, 25-32. Retrieved: 10/14/05, from ACM Portal.
- McCarley, J. S. (1999). Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association For Computational Linguistics on Computational Linguistics* (College Park, Maryland, June 20 - 26, 1999). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 208-214. Retrieved November 10, 2005, from <http://acl.ldc.upenn.edu/P/P99/P99-1027.pdf>.
- McNamee, P. and 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* (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 159-166. Retrieved 11/22/05, from ACM Portal.
- McNamee, P. and Myfield, J. (2004). Corss-language retrieval using HAIRCUT for CLEF (2004). In *Working Notes for the CLEF 2004 Workshop* (Bath, United Kingdom, September, 2004). CLEF-(2004). Retrieved November 23, 2005, from http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/04.pdf.
- Nie, J., and Cai, J. (2001). Filtering noisy parallel corpora of Web pages. In *IEEE Symposium on Natural Language Processing and Knowledge Engineering* (Tucson, AZ, October), 453-458.
- Nie, J., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Berkeley, California, United States, August 15 - 19, 1999). SIGIR '99. ACM Press, New York, NY, 74-81. Retrieved: 10/17/05, from ACM Portal.
- Oard, D. W. (1998). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. In *Proceedings of the Third Conference of the Association For Machine Translation in the Americas on Machine Translation and the information Soup* (October 28 - 31, 1998). D. Farwell, L. Gerber, and E. H. Hovy, Eds. Lecture Notes In Computer Science, vol. 1529. Springer-Verlag, London, 472-483. Retrieved November 23, 2005, from <http://citeseer.ist.psu.edu/oard98comparative.html>.
- Picchi, E., Peters, C. (1998). Cross-language information retrieval: A system for comparable corpus querying. In G. Grefenstette ed. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, 81-92.
- 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* (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, 55-63. Retrieved: 10/4/05, from ACM Portal.
- Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001). Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval* 4, 3-4 (Sep. 2001), 209-230. Retrieved November 21, 2005, from http://www.info.uta.fi/tutkimus/fire/archive/dictionary_based.pdf.
- Pirkola, A., Cosijn, E., Bothma, T., Nel, J. (2002). Cross-lingual information access in indigenous languages: a case study in Zulu language. In *Emerging frameworks and methods, Proceedings of the Fourth International*

- Conference on Conceptions of Library and Information Science, CoLIS4, Seattle, USA, 21 - 25 July 2002.*
Retrieved November 29, 2005, from <http://ucdata.berkeley.edu:7101/sigir-2002/sigir2002CLIR-10-pirkola.pdf>.
- Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, 275-281. Retrieved 11/10/05, from ACM Portal.
- Resnik, P. and Smith, N. A. 2003. The Web as a parallel corpus. *Computational Linguistic*, vol. 29, no. 3, 349-380. Retrieved October 14, 2005, from <http://nlp.cs.jhu.edu/~nasmith/webascorpus.pdf>.
- Rogati, M. and Yang, Y. (2004). Resource selection for domain-specific cross-lingual IR. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 154-161. Retrieved on 10/17/05, from ACM Portal.
- Sheridan, P. and Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Zurich, Switzerland, August 18 - 22, 1996). SIGIR '96. ACM Press, New York, NY, 58-65. Retrieved on 10/4/05, from ACM Portal.
- Xu, J. and Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*. 18, 1 (Jan. 2000), 79-112. Retrieved: 11/4/05, from ACM Portal.
- Xu, J., and Weischedel, R. (2000). Cross-lingual information retrieval using Hidden Markov models. In *Proceeding of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong, October 7-8, (2000). Retrieved November 23, 2005, from <http://acl.ldc.upenn.edu/W/W00/W00-1312.pdf>.
- Xu, J. and Weischedel, R. (2005). Empirical studies on the impact of lexical resources on CLIR performance. *Information Processing and Management*, vol. 41, no. 3, 475-487. Retrieved 11/17/05, from Elsevier.
- Xu, J., Weischedel, R., and Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 105-110. Retrieved: 10/17/05, from ACM Portal.
- Zhang, Y. and Vines, P. (2004). Using the web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 162-169.