Automated query translation in cross-language information retrieval:

A study in tools, techniques and methodologies

Chad Morris
INFO 522: Information Access & Resources
Linda Marion, Instructor
18 March 2010

Introduction

The field of cross-language information retrieval (CLIR) – a subset of the field of information retrieval – deals specifically with the retrieval of documents in one language based on a query formulated in a different language. The language of the document collection is referred to as the target language and the language of the query as the source language. There are two main forms of CLIR: bilingual retrieval, in which there is one source language and one target language, and multilingual retrieval, in which there may be multiple source languages, multiple target languages or both.

Overall, CLIR is a relatively new but growing and increasingly important subset of research within the broader field of information retrieval. Study of the results of this research is important not only for CLIR researchers and system developers but also for librarians and other information professionals who use or may use these systems in their everyday research activities. Study of current research in the field provides valuable background information that can help users more effectively utilize CLIR systems and, more importantly, understand their limitations.

That being said, the research literature within the field of CLIR is extensively varied and prolific. Consequently, the scope of the present bibliography has been limited both in subject matter and by publication date in order to provide a more comprehensive picture of the current research in a specific CLIR model, namely fully automated query translation. In order to provide a complete picture of the contemporary developments in this area, research literature will be presented that deals with the tools, techniques and methodologies used in this model. These primarily include those articles presenting research on translation methodologies, morphological normalization and

corpora acquisition. All articles presented were published between 2007 and the first quarter of 2010.

The remainder of this bibliography is organized as follows. The literature review section contains a more detailed description of CLIR, with particular attention to those topics covered by the literature presented in this bibliography. Citation information as well as abstracts and critical annotations for the research literature reviewed in this bibliography is presented next. Following the bibliography, the selected background reading section includes citations, with limited explanations of content, for resources that contain information that might prove useful for those readers without a background in CLIR. Lastly, the conclusion section provides a personal evaluation of the process of compiling this bibliography.

Literature Review

There are two major models for CLIR systems. The first model depends on translation of the target document collection into the source language, after which more monolingual-focused information retrieval techniques and methodologies are applied to retrieve relevant documents. This model is computationally intensive and, given the relative infancy of machine-translation technologies and the complexities of natural language, is prone to imprecision. The second CLIR model focuses on translation of the query from source language to target language before retrieval takes place. This model is known generally by the term query translation.

Given the relative brevity of queries and their reasonably simple structure when compared to natural language, their translation is much more amenable to machine-translation. Consequently, the query translation model is by far the most common in the contemporary literature and, as such, is the focus of this bibliography. In most bilingual systems, queries are translated directly

from the source language to the target language; however, in some multilingual systems and in cases where few resources exist for direct translation between the source and target language an approach known as transitive translation is often used. In this approach, source language queries are first translated into a pivot language and then from that pivot language into the target language. There is still relatively little research being done with these systems; however, Lehtokangas, Keskustalo and Jarvelin (2008) present one such transitive translation methodology that utilizes a statistical query expansion model known as pseudo relevance feedback to increase recall and precision.

In general, the field of query translation can be further subdivided based on the main translation methodology utilized. Primary reliance on a machine-readable bilingual dictionary is one such methodology, often referred to as dictionary translation. In this methodology, source language query terms entered by the user are replaced by their target language equivalents as stated in the bilingual dictionary used. Another common query translation approach, known as corpus-based translation, is based on translation information derived statistically from bilingual or multilingual document collections, or corpora.

For the purposes of CLIR, these corpora must be related in some way. If the collections are direct translations of one another, such as the publications put out by the United Nations, they are known as parallel corpora. Existence of these officially multilingual collections is rather rare; so, CLIR systems most often employ comparable corpora – collections that share a lot of vocabulary because of related content but are not direct translations of one another. Even professionally prepared comparable corpora are often hard to come by though, and thus some recent research has been devoted to methods for acquiring and aligning reliable corpora for use in CLIR applications. Talvensaari et al. (2007a and 2007b) present methods for aligning fairly loosely

related document collections for effective use as comparable corpora. Methods of mining reliable domain-specific comparable corpora from the Internet have also been proposed (Li et al., 2009; Talvensaari et al., 2008).

Dictionary-based translation and corpus-based translation are the two primary translation technologies employed in modern query translation CLIR systems. Most effective CLIR systems, however, combine these primary translation technologies with each other and other techniques in order make up for their inherent pitfalls as standalone systems. For example, use of dictionary-based translation alone is limited by the scope of available bilingual dictionaries. Most bilingual dictionaries are general purpose in nature, excluding most proper nouns and technical terms. In CLIR, these words that are outside of the scope of the main translation tool being utilized are known as out-of-vocabulary (OOV) words. In pure dictionary translation, OOV words are either passed through as is or are removed from the query altogether. Given that these same words are often primary query terms, both of these courses of action have a seriously negative impact on both precision and recall.

Several methods have been developed by researchers to help deal with OOV words in order to increase the precision and recall of CLIR systems. While some of these methods are based on direct user interaction and are thus out of the scope of this bibliography, most rely heavily on statistical techniques, singularly or in combination, for identifying translation or transliteration alternatives for these terms. Some have proposed corpus-based unification of semantically related terms in both the source and target languages (Li et al., 2009). While others have proposed frequency-based models that choose from translation alternatives created by morphological transformation rules mined from bilingual dictionaries (Pirkola et al., 2008).

Nevertheless, n-gram based approaches are by far the most common in contemporary literature. The basic idea behind the approach is that semantically similar words can be matched by statistical analysis of their character substrings, called grams. The models based in this type of approach vary greatly both in their statistical matching methods and in the characterization of their grams. One such gram characterization variance has been termed skip-gramming, or s-gramming, in which grams are developed not only from adjacent characters but also from non-adjacent characters (Airio and Nuernberger, 2009). Others researchers have proposed n-gram techniques that are enhanced by techniques such as part-of-speech disambiguation (Bellaachia and Amor-Tijani, 2008).

N-gram techniques are also employed in another important aspect of CLIR methodology: morphological normalization. The term morphological normalization refers to the process of reducing word form variation by transforming all inflected word forms to their root, stem or lemma, thus conflating morphological variants into a single representative form. Normalization usually occurs at two distinct steps in CLIR. The first is done when creating the target index. In this case, normalization increases retrieval performance by unifying semantically related terms. The second step at which normalization is important is before query translation. Query words are normalized in order to facilitate translation by dictionaries, which often contain only linguistically valid root word forms as entries not inflected forms, or by other statistical translation techniques.

The two most common normalization techniques employed in information retrieval are lemmatization and stemming. Stemming reduces inflected word forms to a stem or base form that may or may not be the morphological root, utilizing purely statistical methods and ignoring contextual information. Lemmatization, on the other hand, uses contextual information in order

to reduce inflected forms to their morphological root, or lemma, utilizing both statistical methods and linguistic knowledge. While stemming is usually computationally easier to implement, lemmatization is a much more effective normalization method, especially for dictionary-based translation methods. The reason for this is that lemmas are most often the word form used in dictionary entries. Unfortunately, effective lemmatization is still difficult for morphologically complex language such as Semitic and Slavic languages and for languages that have been the subject of very little research. For this reason, most normalization techniques inevitably fall somewhere between stemming and lemmatization.

While some studies have been done with non-normalized indexes and queries, the results indicate that their performance is well below that of their normalized counterparts (Airio and Kettunen, 2009). Consequently, research into effective morphological normalization methods is still important in CLIR. As stated earlier, use of n-gram-based techniques is common in current normalization research, especially for morphologically complex languages. Recently, a lot of attention has been paid to normalization techniques for Arabic, which has historically been hard to normalize because of its extensive use of affixes and root transformation rules. Hmeidi et al. (2010) have found that digrams – substrings composed of two characters – are successful at determining Arabic roots of varying lengths. Additionally, an n-gram technique that takes into account the order of character substrings, an element usually ignored in other n-gram methods, has proven quite effective with conflation of Arabic terms (Ahmed and Nuernberger, 2009).

While popular, n-gram-based techniques are not the only ones employed. Snajder, Basic and Tadic (2008), for example, present an automated lexicon-based morphological normalization process that is based in an original morphology representation formalism. Moreover, just as OOV words cause problems in translation methodologies, so too do they cause problems in

normalization depending on the method being used. Based on the fact that most OOV words are nouns, the methodology introduced by Khaltar and Fujii (2009) holds that words should be normalized using different pathways depending on their part-of-speech. Further distinction and differential treatment of loanwords in this methodology also helped increase successful normalization of words typically considered OOV.

Bibliography

Ahmed, F., & Nuernberger, A. (2009). Evaluation of N-gram conflation approaches for Arabic text retrieval. *Journal of the American Society for Information Science and Technology, 60*(7), 1448-1465.

Abstract: In this paper we present a language-independent approach for conflation that does not depend on predefined rules or prior knowledge of the target language. The proposed unsupervised method is based on an enhancement of the pure n-gram model that can group related words based on various string-similarity measures, while restricting the search to specific locations of the target word by taking into account the order of n-grams. We show that the method is effective to achieve high score similarities for all word-form variations and reduces the ambiguity, i.e., obtains a higher precision and recall, compared to pure n-gram-based approaches for English, Portuguese, and Arabic. The proposed method is especially suited for conflation approaches in Arabic, since Arabic is a highly inflectional language. Therefore, we present in addition an adaptive user interface for Arabic text retrieval called "araSearch". araSearch serves as a metasearch interface to existing search engines. The system is able to extend a query using the proposed conflation approach such that additional results for relevant subwords can be found automatically.

Annotation: The revision to the pure n-gram approach currently used in information retrieval that is presented in the article seems to perform better, with higher recall and precision, than its typically used counterpart. The improvements in precision shown in the experiments presented in this article speaks to the proposed technique's ability to reduce the ambiguity introduced into queries by pure n-gram approaches and, because it is based on n-grams and has no language-dependent rules, it can easily be applied without prior knowledge of the target language. Additionally, the fact that it has been shown to work well on Arabic means that it is likely applicable to other morphologically complex languages, which notoriously provide difficulty in CLIR.

Search strategy: After looking at the results of my initial searches in DIALOG, I determined which journals contained the largest number of relevant articles. Because I wanted to look at the most recent developments in CLIR tools and techniques, I decided to browse through the most recent issues of these journals in order to make sure that I caught relevant articles that may have been missed by my other search strategies.

Database: Wiley InterScience

Method of searching: Browsing

Search string: Browsing the contents of the *Journal of the American Society for Information Science and Technology* for the years 2008 to the present.

Airio, E., & Kettunen, K. (2009). Does dictionary based bilingual retrieval work in a non-normalized index? *Information Processing & Management, 45*(6), 703-713.

Abstract: Many operational IR indexes are non-normalized, i.e. no lemmatization or stemming techniques, etc. have been employed in indexing. This poses a challenge for dictionary-based cross-language retrieval (CLIR), because translations are mostly lemmas. In this study, we face the challenge of dictionary-based CLIR in a non-normalized index. We test two optional approaches: FCG (Frequent Case Generation) and s-gramming. The idea of FCG is to automatically generate the most frequent inflected forms for a given lemma. FCG has been tested in monolingual retrieval and has been shown to be a good method for inflected retrieval, especially for highly inflected languages. S-gramming is an approximate string matching technique (an extension of n-gramming). The language pairs in our tests were English-Finnish, English-Swedish, Swedish-Finnish and Finnish-Swedish. Both our approaches performed quite well, but the results varied depending on the language pair. S-gramming and FCG performed quite equally in all the other language pairs except Finnish-Swedish, where s-gramming outperformed FCG.

Annotation: This article presents some interesting and original work for the field of CLIR, that of query performance in a non-normalized index. The study is well structured and inclusive, covering a range of language pairings and retrieval techniques, including various combinations of frequent case generation and s-gramming. While the various techniques presented outperformed raw runs of translated queries against a non-normalized index, all runs failed to perform at the level of lemmatized runs against a lemmatized index. Nonetheless, the results are promising as a first foray into how to make CLIR viable with a non-normalized index, a case that could be common for target collections in which no normalization options exist for the language in question.

Search strategy: After looking at the results of my initial searches in DIALOG, I determined which journals contained the largest number of relevant articles. Because I wanted to look at the most recent developments in CLIR tools and techniques, I decided to browse through the most recent issues of these journals in order to make sure that I caught relevant articles that may have been missed by my other search strategies.

Database: ScienceDirect

Method of searching: Browsing

Search string: Browsing the contents of *Information Processing and Management* for the years 2008 to the present.

Bellaachia, A., & Amor-Tijani, G. (2008). Proper nouns in English-Arabic cross language information retrieval. *Journal of the American Society for Information Science and Technology, 59*(12), 1925-1932.

Abstract: Out of vocabulary words, mostly proper nouns and technical terms, are one main source of performance degradation in Cross Language Information Retrieval (CLIR) systems. Those are words not found in the dictionary. Bilingual dictionaries in general do not cover most proper nouns, which are usually primary keys in the query. As they are spelling variants of each other in most languages, using an approximate string matching technique against the target database index is the common approach taken to find the target language correspondents of the original query key. N-gram technique proved to be the most effective among other string matching techniques. The issue arises when the languages dealt with have different alphabets. Transliteration is then applied based on phonetic similarities between the languages involved. In this study, both transliteration and the n-gram technique are combined to generate possible

transliterations in an English-Arabic CUR system. We refer to this technique as Transliteration N-Gram (TNG). We further enhance TNG by applying Part Of Speech disambiguation on the set of transliterations so that words with a similar spelling, but a different meaning, are excluded. Experimental results show that TNG gives promising results, and enhanced TNG further improves performance.

Annotation: The authors of this article present a viable approach to dealing with OOV words arising from dictionary translation methods when the orthographies of the source and target languages are different. Previous work in this area has relied heavily on statistical transliteration techniques to produce a set of possible transliterations for a source language OOV term. The approach presented here, however, focuses on manipulation of a single transliterated form of an OOV query term. The proposed two step process first applies n-gram techniques to a transliterated OOV query term and then further disambiguates the resulting variants from the target document collection based on part-of-speech analysis.

The testing of an exhaustive array of character combinations in the n-gram technique employed in this study is noteworthy. The approach proposed in this article shows promising results for enhancing retrieval of documents that contain transliterated forms of proper nouns, which are often primary query terms. While the study is focused on English-Arabic systems, the approach may be applicable and well-suited to other orthography pairings.

Search strategy: I chose INSPEC because it indexes science and engineering literature with abstracts. The subject coverage of this particular database is helpful because much of literature on CLIR appears in scientific and engineering journals. This was during my early searching and so most of my searching was performed with keywords with some term expansion for clarity.

Database: INSPEC [DIALOG File 2]

Method of searching: Keyword search

Search string: S (CLIR OR CROSS LANGUAGE INFORMATION RETRIEVAL) AND DT=JOURNAL PAPER

sort by publication year

Hmeidi, I. I., Al-Shalabi, R. F., Al-Taani, A. T., Najadat, H., & Al-Hazaimeh, S. A. (2010). A novel approach to the extraction of roots from Arabic words using bigrams. *Journal of the American Society for Information Science and Technology, 61*(3), 583-591.

Abstract: Root extraction is one of the most important topics in information retrieval (IR), natural language processing (NLP), text summarization, and many other important fields. In the last two decades, several algorithms have been proposed to extract Arabic roots. Most of these algorithms dealt with triliteral roots only, and some with fixed length words only. In this study, a novel approach to the extraction of roots from Arabic words using bigrams is proposed. Two similarity measures are used, the dissimilarity measure called the "Manhattan distance," and Dice's measure of similarity. The proposed algorithm is tested on the Holy Qu'ran and on a corpus of 242 abstracts from the Proceedings of the Saudi Arabian National Computer Conferences. The two files used contain a wide range of data: the Holy Qu'ran contains most of the ancient Arabic words while the other file contains some modern Arabic words and some words borrowed from foreign languages in addition to the original Arabic words. The results of this study showed that combining N-grams with the Dice measure gives better results than using the Manhattan distance measure.

Annotation: Effective lemmatization or stemming of index terms is a necessary and important step, known as morphological normalization, in almost all query translation approaches. This process, however, can be especially difficult for some morphologically complex languages. Semitic languages, of which Arabic is one, are a case for which this is particularly true. In these languages, consonantal roots are the canonical form, or lemma, of verbs and nouns; therefore, their extraction is essential to effective query translation, especially dictionary-based translation. This article presents an effective approach for finding the roots of Arabic words based on multiple n-gram techniques combined with statistical similarity measures. Similar approaches could be built upon to improve lemmatization of Arabic and other Semitic languages.

Search strategy: After looking at the results of my initial searches in DIALOG, I determined which journals contained the largest number of relevant articles. Because I wanted to look at the most recent developments in CLIR tools and techniques, I decided to browse through the most recent issues of these journals in order to make sure that I caught relevant articles that may have been missed by my other search strategies.

Database: Wiley InterScience

Method of searching: Browsing

Search string: Browsing the contents of the *Journal of the American Society for Information Science and Technology* for the years 2008 to the present.

Khaltar, B., & Fujii, A. (2009). A lemmatization method for Mongolian and its application to indexing for information retrieval. *Information Processing & Management, 45*(4), 438-451.

Abstract: In Mongolian, two different alphabets are used, Cyrillic and Mongolian. In this paper, we focus solely on the Mongolian language using the Cyrillic alphabet, in which a content word can be inflected when concatenated with one or more suffixes. Identifying the original form of content words is crucial for natural language processing and information retrieval. We propose a lemmatization method for Mongolian. The advantage of our lemmatization method is that it does not rely on noun dictionaries, enabling us to lemmatize out-of-dictionary words. We also apply our method to indexing for information retrieval. We use newspaper articles and technical abstracts in experiments that show the effectiveness of our method. Our research is the first significant exploration of the effectiveness of lemmatization for information retrieval in Mongolian.

Annotation: The lemmatization method presented in this article is important for two main reasons: 1) it is designed for Mongolian, a language which has historically had very few morphological normalization studies and 2) it differentiates words for processing based on parts of speech. By lemmatizing nouns differently than verbs, this method was better able to handle words that would be normally be considered OOV by standard dictionary based methods, namely loanwords. This is promising in that it demonstrates that lemmatization methods that provide different pathways based on part-of-speech disambiguation may obtain better results than those that provide only a single pathway, thus increasing information retrieval performance.

Search strategy: After looking at the results of my initial searches in DIALOG, I determined which journals contained the largest number of relevant articles. Because I wanted to look at the most recent developments in CLIR tools and techniques, I decided to browse through the most recent issues of these journals in order to make sure that I caught relevant articles that may have been missed by my other search strategies.

Database: ScienceDirect

Method of searching: Browsing

Search string: Browsing the contents of *Information Processing and Management* for the years 2008 to the present.

Lehtokangas, R., Keskustalo, H., & Jarvelin, K. (2008). Experiments with transitive dictionary translation and pseudo-relevance feedback using graded relevance assessments. *Journal of the American Society for Information Science and Technology, 59*(3), 476-488.

Abstract: In this article, the authors present evaluation results for transitive dictionary-based cross-language information retrieval (CLIR) using graded relevance assessments in a best match retrieval environment. A text database containing newspaper articles and a related set of 35 search topics were used in the tests. Source language topics (in English, German, and Swedish) were automatically translated into the target language (Finnish) via an intermediate (or pivot) language. Effectiveness of the transitively translated queries was compared to that of the directly translated and monolingual Finnish queries. Pseudo-relevance feedback (PRF) was also used to expand the original transitive target queries. Cross-language information retrieval performance was evaluated on three relevance thresholds: stringent, regular, and liberal. The transitive translations performed well achieving, on the average, 85-93% of the direct translation performance, and 66-72% of monolingual performance. Moreover, PRF was successful in raising the performance of transitive translation routes in absolute terms as well as in relation to monolingual and direct translation performance applying PRF.

Annotation: Transitive query translation, or the translation of a query in a source language into the target language by means of an intermediary language known as a pivot language, is an

important approach in CLIR for two specific cases. The first case in which this approach is crucial is the need for information retrieval across languages for which there are no or very few direct translations resources. The second case is in systems that need to handle multiple source and target languages. Most studies in transitive query translation are performed using dictionary-based methods. This study is no exception; however, it is unique and its findings important because of two factors: 1) its use of a graded relevance assessment model, which is more aligned with typical user behavior than the characteristic binary relevance model used in many laboratory-based IR studies, and 2) its exploration of a pseudo relevance feedback mechanism for query expansion to improve base transitive translation performance.

Search strategy: I chose INSPEC because it indexes science and engineering literature with abstracts. The subject coverage of this particular database is helpful because much of literature on CLIR appears in scientific and engineering journals. This was during my early searching and so most of my searching was performed with keywords with some term expansion for clarity.

Database: INSPEC [DIALOG File 2]

Method of searching: Keyword search

Search string: S (CLIR OR CROSS LANGUAGE INFORMATION RETRIEVAL) AND DT=JOURNAL PAPER

sort by publication year

Li, Q., Chen, Y. P., Myaeng, S., Jin, Y., & Kang, B. (2009). Concept unification of terms in different languages via Web mining for information retrieval. *Information Processing & Management, 45*(2), 246-262.

Abstract: For historical and cultural reasons, English phrases, especially proper nouns and new words, frequently appear in Web pages written primarily in East Asian languages such as Chinese, Korean, and Japanese. Although such English terms and their equivalences in these East Asian languages refer to the same concept, they are often erroneously treated as independent index units in traditional Information Retrieval (IR). This paper describes the degree to which the problem arises in IR and proposes a novel technique to solve it. Our method first extracts English terms from native Web documents in an East Asian language, and then unifies the extracted terms and their equivalences in the native language as one index unit. For Cross-Language Information Retrieval (CLIR), one of the major hindrances to achieving retrieval performance at the level of Mono-Lingual Information Retrieval (MLIR) is the translation of terms in search queries which cannot be found in a bilingual dictionary. The Web mining approach proposed in this paper for concept unification of terms in different languages can also be applied to solve this well-known challenge in CLIR. Experimental results based on NTCIR and KT-Set test collections show that the high translation precision of our approach greatly improves performance of both Mono-Lingual and Cross-Language Information Retrieval.

Annotation: The authors of this article hold that query translation-based CLIR performance can be enhanced by concept unification, or the unifying of semantically identical terms in the source and target language, to handle OOV words that might otherwise be ignored in a traditional dictionary-based query translation approach. In particular, the focus is on target language documents that contain embedded source language terms. This is important because of the increasing prevalence of language mixing, especially in East Asian language documents of a highly technical nature. The concept unification approach presented in this article relies on a

statistically-, semantically- and phonetically-based model trained from a small corpus of Web document snippets returned by an intelligent search engine, such as Google, based on a domain specific word list compiled from reliable domain-relevant mixed language Web documents. The results of this study demonstrate that this approach could prove to be a highly portable and adaptable method for handling OOV words in rapidly changing domains for language pairs where language mixing is prevalent.

Search strategy: I chose INSPEC because it indexes science and engineering literature with abstracts. The subject coverage of this particular database is helpful because much of literature on CLIR appears in scientific and engineering journals. This was during my early searching and so most of my searching was performed with keywords with some term expansion for clarity.

Database: INSPEC [DIALOG File 2]

Method of searching: Keyword search

Search string: S (CLIR OR CROSS LANGUAGE INFORMATION RETRIEVAL) AND DT=JOURNAL PAPER

sort by publication year

Pirkola, A., Toivonen, J., Keskustalo, H., & Jarvelin, K. (2008). Frequency-based identification of correct translation equivalents (FITE) obtained through transformation rules. *ACM Transactions on Information Systems, 26*(1), 2.

Abstract: We devised a novel statistical technique for the identification of the translation equivalents of source words obtained by transformation rule based translation (TRT). The effectiveness of the technique called frequency-based identification of translation equivalents

(FITE) was tested using biological and medical cross-lingual spelling variants and out-of-vocabulary (OOV) words in Spanish-English and Finnish-English TRT. The results showed that, depending on the source language and frequency corpus, FITE-TRT (the identification of translation equivalents from TRT's translation set by means of the FITE technique) may achieve high translation recall. In the case of the Web as the frequency corpus, translation recall was 89.20%-91.0% for Spanish-English FITE-TRT. For both language pairs FITE-TRT achieved high translation precision: 95.0%-98.8%. The technique also reliably identified native source language words: source words that cannot be correctly translated by TRT. Dictionary-based CLIR augmented with FITE-TRT performed substantially better than basic dictionary-based CLIR where OOV keys were kept intact. FITE-TRT with Web document frequencies was the best technique among several fuzzy translation/matching approaches tested in cross-language retrieval experiments. We also discuss the application of FITE-TRT in the automatic construction of multilingual dictionaries.

Annotation: This article presents a refinement in an earlier fuzzy logic translation technique developed by the authors known as transformation rules (TRT). More specifically, the refined technique is meant to determine the correction translation from the list of possible translations produced by TRT based on frequency data collected from a domain relevant corpus. The technique shows promise, upon further investigation and refinement, as a viable tool for translating OOV words resulting from more typical dictionary-based query translation methods. Additionally, as the authors note, the method presented may be useful as a tool to automatically create multilingual dictionaries of technical terms and proper nouns for use in CLIR applications, a much cheaper and less labor-intensive alternative to similar hand-created translation resources.

Search strategy: Ari Pirkola was an author name that came up time and again in my readings for this bibliography. Because of this, I thought to search for the most recent works by this author.

Database: Web of Science

Method of searching: Author search

Search string: Author=(pirkola a*) with a Timespan=Latest 5 years

Snajder, J., Basic, B. D., & Tadic, M. (2008). Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing & Management, 44*(5), 1720-1731.

Abstract: Due to natural language morphology, words can take on various morphological forms. Morphological normalisation - often used in information retrieval and text mining systems - conflates morphological variants of a word to a single representative form. In this paper, we describe an approach to lexicon-based inflectional normalisation. This approach is in between stemming and lemmatisation, and is suitable for morphological normalisation of inflectionally complex languages. To eliminate the immense effort required to compile the lexicon by hand, we focus on the problem of acquiring automatically an inflectional morphological lexicon from raw corpora. We propose a convenient and highly expressive morphology representation formalism on which the acquisition procedure is based. Our approach is applied to the morphologically complex Croatian language, but it should be equally applicable to other languages of similar morphological complexity. Experimental results show that Our approach can be used to acquire a lexicon whose linguistic quality allows for rather good normalisation performance.

Annotation: The authors of this paper present a lexicon-based automated morphological normalization process that is based in an original morphology representation formalism. The automated building of this lexicon from an existing corpus is also unique. Moreover, the formalism presented is applicable to other languages and is especially suited to those languages whose morphological complexity stems from complex inflection rules. The study presented in this article, compares the proposed approach against typical lemmatization based on an expertly prepared morphological lexicon; however, comparison to stemming – the second of the two most typical normalization approaches used in CLIR – and testing within an information retrieval context are important measures that are not present in the current study. Nevertheless, the approach shows promise and, in the future, may be extremely useful for normalization in CLIR applications utilizing languages for which no formal morphological lexicon has been produced by linguistic experts.

Search strategy: After looking at the results of my initial searches in DIALOG, I determined which journals contained the largest number of relevant articles. Because I wanted to look at the most recent developments in CLIR tools and techniques, I decided to browse through the most recent issues of these journals in order to make sure that I caught relevant articles that may have been missed by my other search strategies.

Database: ScienceDirect

Method of searching: Browsing

Search string: Browsing the contents of *Information Processing and Management* for the years 2008 to the present.

Talvensaari, T., Juhola, M., Laurikkala, J., & Jarvelin, K. (2007a). Corpus-based cross-language information retrieval in retrieval of highly relevant documents. *Journal of the American Society for Information Science and Technology, 58*(3), 322-334.

Abstract: Information retrieval systems' ability to retrieve highly relevant documents has become more and more important in the age of extremely large collections, such as the World Wide Web (WWW). The authors' aim was to find out how corpus-based cross-language information retrieval (CLIR) manages in retrieving highly relevant documents. They created a Finnish-Swedish comparable corpus from two loosely related document collections and used it as a source of knowledge for query translation. Finnish test queries were translated into Swedish and run against a Swedish test collection. Graded relevance assessments were used in evaluating the results and three relevance criterion levels-liberal, regular, and stringent-were applied. The runs were also evaluated with generalized recall and precision, which weight the retrieved documents according to their relevance level. The performance of the Comparable Corpus Translation system (COCOT) was compared to that of a dictionary-based query translation program; the two translation methods were also combined. The results indicate that corpus-based CUR performs particularly well with highly relevant documents. In average precision, COCOT even matched the monolingual baseline on the highest relevance level. The performance of the different query translation methods was further analyzed by finding out reasons for poor rankings of highly relevant documents.

Annotation: When read in conjunction with Talvensaari et al. (2007b), this article provides a complete CLIR methodology using query translation based on document-aligned comparable corpora drawn from disparate origin. This article is particularly useful for two reasons: 1) its testing of the proposed pivoted vector length normalization against an extensive set of previously

developed translation methods and 2) its use of a graded relevance assessment model, which is more aligned with typical user behavior than the characteristic binary relevance model used in many laboratory-based IR studies.

Search strategy: I chose INSPEC because it indexes science and engineering literature with abstracts. The subject coverage of this particular database is helpful because much of literature on CLIR appears in scientific and engineering journals. This was during my early searching and so most of my searching was performed with keywords with some term expansion for clarity.

Database: INSPEC [DIALOG File 2]

Method of searching: Keyword search

Search string: S (CLIR OR CROSS LANGUAGE INFORMATION RETRIEVAL) AND DT=JOURNAL PAPER

sort by publication year

Talvensaari, T., Laurikkala, J., Jarvelin, K., Juhola, M., & Keskustalo, H. (2007b). Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems, 25*(1), 4.

Abstract: We present a method for creating a comparable text corpus from two document collections in different languages. The collections can be very different in origin. In this study, we build a comparable corpus from articles by a Swedish news agency and a U.S. newspaper. The keys with best resolution power were extracted from the documents of one collection, the source collection, by using the relative average term frequency (RATF) value. The keys were translated into the language of the other collection, the target collection, with a dictionary-based

query translation program. The translated queries were run against the target collection and an alignment pair was made if the retrieved documents matched given date and similarity score criteria. The resulting comparable collection was used as a similarity thesaurus to translate queries along with a dictionary-based translator. The combined approaches outperformed translation schemes where dictionary-based translation or corpus translation was used alone.

Annotation: When read in conjunction with Talvensaari et al. (2007a), this article provides a complete CLIR methodology using query translation based on document-aligned comparable corpora drawn from disparate origin. This article goes into the details of creating the document-aligned comparable corpora. The proposed method is important because it does not presuppose content knowledge of the monolingual corpora utilized to create the final collection and, in the proposed study, this corpus shows promise when used in creating a similarity thesaurus to handle OOV words.

Search strategy: I performed some forward citation searching using Web of Science for articles that I had already decided on including in the bibliography to see if any more recent research existed along the same vein.

Database: Web of Science

Method of searching: Citation search

Search string: Forward citation searching from Talvensaari et al. (2007a).

Talvensaari, T., Pirkola, A., Jarvelin, K., Juhola, M., & Laurikkala, J. (2008). Focused Web crawling in the acquisition of comparable corpora. *Information Retrieval, 11*(5), 427-445.

Abstract: Cross-Language Information Retrieval (CLIR) resources, such as dictionaries and parallel corpora, are scarce for special domains. Obtaining comparable corpora automatically for such domains could be an answer to this problem. The Web, with its vast volumes of data, offers a natural source for this. We experimented with focused crawling as a means to acquire comparable corpora in the genomics domain. The acquired corpora were used to statistically translate domain-specific words. The same words were also translated using a high-quality, but non-genomics-related parallel corpus, which fared considerably worse. We also evaluated our system with standard information retrieval (IR) experiments, combining statistical translation using the Web corpora with dictionary-based translation. The results showed improvement over pure dictionary-based translation. Therefore, mining the Web for comparable corpora seems promising.

Annotation: This article introduces a highly adaptive and portable approach to generating paragraph-aligned comparable corpora from the Web using a small number of known reliable domain-relevant source websites. Building on their earlier work on automatic corpora generation (Talvensaari et al. 2007b), the authors present an approach that is specifically designed for acquisition of domain-specific corpora. Use of the resulting corpora shows promising results in enhancing dictionary-based query translation methods when dealing with rapidly changing domains, whose queries contain a high frequency of domain-specific vocabulary that would otherwise remain as OOV words using more traditional CLIR methods based on pure dictionary translation or established research corpora.

Search strategy: I chose INSPEC because it indexes science and engineering literature with abstracts. The subject coverage of this particular database is helpful because much of literature

on CLIR appears in scientific and engineering journals. This was during my early searching and so most of my searching was performed with keywords with some term expansion for clarity.

Database: INSPEC [DIALOG File 2]

Method of searching: Keyword search

Search string: S (CLIR OR CROSS LANGUAGE INFORMATION RETRIEVAL) AND DT=JOURNAL PAPER

sort by publication year

Selected Background Reading

Many of the articles presented in this bibliography do a reasonably good job of explaining the background of their particular research; however, they can still be hard to understand without a bit of background on the field in general. Readers who have little or no background in CLIR may find the following materials useful in developing a backdrop from which to understand some of the topics presented in the articles in this bibliography.

The following article provides a good overview of CLIR – its background, general models and implementation.

Oard, D. W., & Diekema, A. R. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology, 33*, 223-256.

The following two documents provide information about dictionary-based translation methodology. They do an especially good job of explaining the process step-by-step and explaining how research in the field of CLIR in general operates.

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R., Pirkola, A., & Jarvelin, K. (2004). Dictionary-based cross-language information retrieval: Learning experiences from CLEF 2000-2002. *Information Retrieval, 7*(1-2), 99-119.

Levow, G. A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing & Management, 41*(3), 523-547.

The following article provides an overview of the development, operation and performance testing of UTACLIR, arguable the most popular dictionary-based translation system utilized in contemporary CLIR research.

Pirkola, A., Hedlund, T., Keskustalo, H., & Jarvelin, K. (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval, 4*(3-4), 209-230.

The following articles introduce the concept of string matching techniques, specifically n-gram-based techniques. As can be seen from the literature presented in this bibliography, n-gram-based techniques are prevalent in CLIR today. Pirkola et al. (2002) presents the specific variation of n-gramming known as skip-gramming, or s-gramming, that is gaining popularity in CLIR research.

Hall, P. A. V., & Dowling, G. R. (1980). Approximate string matching. *Computing Surveys, 12*(4), 381-402.

Pirkola, A., Keskustalo, H., Leppanen, E., Kansala, A. -P., & Jarvelin, K. (2002). Targeted s-gram matching: A novel n-gram matching technique for cross- and monolingual word

form variants. *Information Research, 7*(2) Retrieved from http://informationr.net/ir/7-2/paper126.html

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K., & Jarvelin, K. (2003). Fuzzy translation of cross-lingual spelling variants. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 345-352.

Salton and Buckley (1988) present the term frequency-inverse document frequency (*tf · idf*) formula for determining term similarity. This formula is widely used and manipulated in CLIR, especially for corpus-based translation approaches. Singhal, Buckley and Mitra (1996) introduced a modification to this formula known as pivoted vector-length normalization that tries to combat the *tf · idf* formula's tendency to favor short documents. This normalization technique is becoming increasingly popular in CLIR research.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 21-29.

The following document was instrumental in pointing out the ineffectiveness of the binary relevance model that dominated laboratory IR experiments at the time. While this model still persists in the literature today, there has been a move towards graded relevance models, as presented and pushed for in this document, in more recent research.

Sourmunen, E.Liberal relevance criteria of TREC—Counting on negligible documents? *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 320-330.

Conclusion

Working on this project has been an enjoyable and educational experience. It has given me the opportunity to take a self-guided crash course in CLIR, with some side forays into set theory, vector math, statistics and linguistics. CLIR is an intriguing field that I would like to continue learning more about and the reading for this project, while lengthy and frustratingly dense at times, was rewarding and pleasant in the end. Moreover, this project has allowed me to flex my searching muscles, to put into real-world application the skills and methods I have learned in class. The requirements of the assignment forced me to explore search techniques that I have never or rarely used before and provided for a thorough introduction to using a variety of databases, with differing interfaces, organizational structures and search capabilities.

Additionally, besides needing the skill for future classes, learning how to construct an annotated bibliography may be critical in a professional sense for me. I have been considering professional research positions as a possible career direction. In this light, knowledge of how to construct an annotated bibliography and experience actually performing the task could be an asset. Furthermore, the associated skills learned from undertaking this exercise – such as learning how to evaluation information resources and how to synthesis information – are broadly applicable to most professions in the field of information science. Overall, I feel that this assignment is a great way to reinforce the knowledge gained in class and develop important professional skills while learning about a topic of personal interest.