RESOLVING AMBIGUITY FOR CROSS-LANGUAGE INFORMATION RETRIEVAL: A DICTIONARY APPROACH

A Dissertation Presented

by

LISA ANNE BALLESTEROS

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2001

Department of Computer Science

RESOLVING AMBIGUITY FOR CROSS-LANGUAGE INFORMATION RETRIEVAL: A DICTIONARY APPROACH

A Dissertation Presented

by

LISA ANNE BALLESTEROS

Approved as to style and content by:

W. B. Croft, Chair

Jamie Callan, Member

Paul R. Cohen, Member

Roger Higgins , Member

James Kurose, Department Chair
Department of Computer Science

To my late grandmother who did not witness this success, but whose constant support and faith in me helped to make it possible.

ACKNOWLEDGMENTS

I'd like to acknowledge the many people whose assistance and support helped to make this study possible. Many thanks to Bruce Croft for his advice throughout this research. Special thanks to Jamie Callan for many helpful discussions and for his friendship and willingness to cheer me on when I was discouraged. Thanks also to the other members of my dissertation committee, Paul Cohen and Roger Higgins, whose valuable comments and discussions helped to improve this dissertation. I am particularly grateful to Paul for his mentorship.

I'd also like to thank all CIIR faculty, students, and staff for their helpful conversations, willingness to make software available, and for providing both technical and administrative support throughout this study.

Finally, I would like to express my gratitude and appreciation to my family who never doubted me for a moment. I am especially grateful for the love and friendship of my husband, David, whose unfailing support, encouragement, and sense of humor helped us both to get through this alternately exciting and maddening roller-coaster ride.

This material is based on work supported by the National Science Foundation, Library of Congress and Department of Commerce under the cooperative agreement # EEC-9209623 and by a NASA GSRP grant, # NGT-70358.

ABSTRACT

RESOLVING AMBIGUITY FOR CROSS-LANGUAGE INFORMATION RETRIEVAL: A DICTIONARY APPROACH

SEPTEMBER 2001

LISA ANNE BALLESTEROS
B.Sc., UNION COLLEGE
M.Sc., UNIVERSITY OF MASSACHUSETTS, AMHERST
Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor W. B. Croft

The global exchange of information has been facilitated by the rapid expansion in the size and use of the Internet, which has led to a large increase in the availability of on-line texts. Expanded international collaboration, the increase in the availability of electronic foreign language texts, the growing number of non-English-speaking users, and the lack of a common language of discourse compels us to to develop *cross-language* information retrieval (CLIR) tools capable of bridging the language barrier. Cross-language retrieval bridges this gap by enabling a person to search in one language and retrieve documents across languages.

There are several goals for the research described herein. The first is to gain a clear understanding of the problems associated with the cross-language task and to develop techniques for addressing them. Empirical work shows that ambiguity and lack of lexical resources are the main hurdles. Second we show that cross-language effectiveness does not depend upon linguistic analysis. We demonstrate how statistical techniques can be used to significantly reduce the effects of ambiguity. We also show that combining these techniques is as effective as or more effective than a reasonable machine translation system. Third, we show that an approach based on multi-lingual dictionaries and statistical analysis can be used as the foundation for a cross-language retrieval architecture that circumvents the problem of limited resources.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

In recent years, the amount of online information from the government, scientific, business, and private sectors has risen dramatically. It is currently estimated that 82% of all World Wide Web (”web”) pages are written in English [2]. However, English is not the the native language of nearly half of all Internet users [54] and the number of non-English-speaking users is growing. The diversity of information sources and the explosive growth of the Internet worldwide are compelling evidence of the need for IR systems that span language boundaries. As people have become more aware of global needs and concerns, multi-national collaboration has increased and people are more interested in data and information collected in other regions of the world. It is becoming increasingly important to find ways to facilitate access to the world's databases and to address cross-language issues that arise with global use and sharing of information.

The Internet environment stands to benefit greatly from *Cross-language Information Retrieval* (CLIR) technology. We are only beginning to understand the ways in which we can exploit this powerful resource in areas such as electronic commerce, advertising and marketing, education, research, banking and finance. Regardless of their area of interest, one task that Internet users must address is how to access the huge amount of data and information that is available to them. This includes finding out what information is available, where it is, and determining whether or not it is useful. This process is time consuming and can be overwhelming. It is made more difficult when important information is written in a language that is not the user's first language. The lack of tools that enable the user to more easily cross language boundaries greatly increases the probability that foreign language sources will be overlooked. Oard [94] discusses studies reporting that researchers failed to find foreign language work that was in fact relevant. One can imagine scenarios from any of the areas mentioned above that would benefit from the ability to locate foreign language information in a timely manner.

The advantages of CLIR technology are not limited to individual users of the Internet. Many business, government, social, and multi-national organizations would also benefit from the ability to perform searches across languages. Groups like the World Bank, for example, collect financial information from around the world. Professional translators are hired full-time to translate a backlog of international law documents whose contents may greatly affect policy decisions. A CLIR system would identify a subset of documents likely to be relevant. The documents could then be translated in a timely fashion. This functionality would also benefit a patent officer searching foreign patent databases or a security officer monitoring foreign news and intelligence sources. Regardless of the task, cross-language retrieval would bridge the language gap thus facilitating information gathering from a wider range of sources and freeing people for more important tasks.

The following section describes general information retrieval concepts and techniques. Section 1.3 discusses issues related to the development of systems for retrieving documents in a language other than English. We then describe the cross-language problem in Section 1.4. Finally, the contributions of this research are outlined in Section 1.5 and an outline of the presentation of this dissertation is given.

1.2 Information Retrieval

The task of an Information Retrieval (IR) system is to estimate the degree to which documents in a collection reflect the information expressed in a user query. IR techniques facilitate this access by focusing on three basic processes. The first process is to develop a representation of text objects (i.e. queries and documents). The second process is comparing query and document representations to retrieve those documents most closely matching an information need. The third process is to evaluate the documents retrieved with the possibility of using the evaluation to improve subsequent searches. This section reviews the issues related to and techniques for generating document representations (*indexing*), briefly discusses retrieval models and describes the INQUERY retrieval system, and finally describes standard techniques including the application of relevance feedback for increasing retrieval effectiveness by improving query representations.

1.2.1 Indexing

We can think of documents in a collection and a searcher's query as objects of the system. Objects should be represented in a manner which facilitates operations on them and which balances the trade-offs between efficiency and cost. Object representations are typically based on the words or vocabulary of a language. This seems to make sense because textual information is conveyed using words. However, the ideal representation should do two things. First, it should contain all and only those words necessary to convey the intended meaning of documents and queries. Second, it should contain words that facilitate the partitioning of documents into those that meet an information need and those that do not. We often choose not to use the entire text of an object as written, so one representation issue is deciding how to define a word. In addition to single words, an indexing language may contain collocations or phrases, features such as proper names, and/or thesaurus classes. The process of building this representation is known as *indexing*. *Index term* is the general name for a word, phrase, or feature used in indexing and the full set of index terms is the *indexing language*.

During text processing to generate an index, two additional steps are often performed: *stopword* removal and *stemming*. The former is a way to limit the words that will be represented given that there are a number of words which occur so frequently that they provide no useful information with respect to content. These non-content-bearing words are referred to as stopwords. Stopwords are words like *an*, *the*, and *because*. Stemming is a technique for conflating morphological variants or words believed to have the same root. One way to stem in Latin-based languages is to remove the suffixes of words with the same stem, the simplest being the removal of *'s*. This is based on the idea that words having the same stem refer to the same concept. Although there will be examples where this assumption will be false, the error rate does not significantly degrade effectiveness [77, 60].

In terms of storage, representing documents via inverted files (or lists) provides us with easy access to document information. In an inverted file, a separate index into the file is created for each word. One can think of the structure as an array of indexed records, where each row corresponds to a document and each column represents an index term. This array is then transposed so that each row specifies the documents containing a particular word. The row will contain a record to hold term information for query processing, for each document that the word occurs in. This information may include, for example, the location(s) of the word in the document in addition to the frequency with which it occurs. Positional information can then be used to identify phrases on the fly, thus saving considerable indexing time and storage space.

1.2.2 Retrieval Models

A retrieval model attempts to model aspects of the retrieval process. It must describe a retrieval mechanism for matching queries and documents, but can also describe a means for reformulating a query to more accurately reflect the information need. We will focus here on the former.

The goal of comparing query and document objects is to allow the system to successfully discriminate between relevant and non-relevant documents by generating some measure of similarity between the document and the query. Furthermore, it is typically desirable that the measure allow the ranking of documents such that those most closely matching the query appear first. However, language is ambiguous and there are many different ways to express a given idea or concept. This presents difficulties for ascertaining two things needed to estimate how likely it is that a document will match a user's need. The first is inferring what information the writer of a document was trying to convey to the reader. For example, is a document containing the word "boxers" about dogs or is it about fighters? The second is inferring what information need a user is trying to express with a query.

In the probabilistic retrieval model [106] the statistical analysis of text is the basis for comparison and is developed from two hypotheses. The first is that the probability that a document will be relevant to a query can measure the extent to which the query concept is treated in the document. In other words, query concepts will appear more often in relevant documents than in non-relevant documents. The second is that the frequency of words referring to a concept measures the extent to which the document content is related to that concept. In fact it has been shown that statistical information (e.g. word frequency)[108] can be used to measure the importance of words and sentences in a document. With this in mind, the belief that a document is relevant to the query is based upon the frequency of each query term in the document text.

The inference net model upon which INQUERY is based gets its foundation from probability theory and is based on Bayesian inference nets [122, 20]. In probabilistic retrieval, the extent to which a document is "about" a particular concept is measured by the number of terms it contains which refer to or describe that concept. The weighting of the index terms with respect to the document, is given a theoretical justification in terms of probabilities. The next section describes INQUERY and the inference net model upon which it is built.

1.2.2.1 The INQUERY System

INQUERY, which has been developed at the University of Massachusetts, is based on a powerful inference net framework and has been demonstrated to be very effective in government evaluations [25, 24, 29]. We use the INQUERY system as a framework for the research in this study.

In the inference net model, queries are considered to be representations of the user's information need, and the information need is considered a complex proposition about the content of a text object. Retrieval is considered a process of inference in which the system estimates the probability that the user's information need is met given a document as "evidence". The basic idea is that retrieval performance can be greatly improved if the content of documents and queries is used to infer the probability that they are related. An important aspect of this model is that it allows for the combination of different types of document information such as representations of text generated via different indexing approaches.

INQUERY utilizes a combination of statistical and Natural Language Processing (NLP) techniques to make inferences about the relatedness of the concepts in a query and objects in a collection. NLP techniques attempt to explicitly represent the syntactic or semantic structure of text and include stemming, part of speech (POS) tagging, and phrase formulation. Statistical techniques, on the other hand, use frequency and co-occurrence information to build text representations. Multiple representations of query and document concepts facilitate the inference mechanism since all representations of those concepts can be regarded as multiple sources of evidence that a document and query are related.

Document and query objects are represented by a directed acyclic graph (DAG) where nodes correspond to content representations that can be considered propositions and edges represent dependence relations between nodes. When a node representing one proposition "causes" or implies another proposition, an edge is directed from the former toward the latter.

Positive evidential support for a representation concept or proposition is expressed by a measure of belief representing the certainty that a proposition is true. The belief in a representation concept is dependent upon the frequency with which the concept occurs in a document and upon the frequency of occurrence in the collection. INQUERY's measure of belief is inversely proportional to the frequency with which a concept occurs in the collection (inverse document frequency or *idf*) and directly proportional to the with-in document frequency (term frequency or *tf*). In other words, the belief that a document implies a query is a function of the frequency of occurrence of query terms in that document and of their frequency of occurrence in the corpus being searched.

A good function for measuring the degree of match between a query and document may not be enough to ensure effective retrieval. Effectiveness is typically measured by *precision* and *recall*, which are the proportion of retrieved documents that are relevant and the proportion of relevant documents retrieved, respectively [107]. A more detailed discussion of evaluation is given in Chapter 4. Given that a user will often not clearly specify an information need, it is useful to generate modified queries which will more accurately reflect that need. The next section describes automatic techniques for improving retrieval effectiveness.

1.2.3 Retrieval Techniques

This section considers techniques for automatically exploiting characteristics of text in order to improve search effectiveness. Given a query and a collection of documents, these techniques focus primarily on reformulating the query to more accurately describe the topic of the search. The first approach is query expansion, in which words related to query terms are added to the query. The second approach is the application of query structure, which aims to reinforce the concepts conveyed by the query. Finally, we describe co-occurrence analysis for identification of semantic relationships between words.

1.2.3.1 Query Expansion

Word mismatch or *vocabulary mismatch* is a common problem in information retrieval. It occurs when the vocabulary chosen by a person to express an information need differs from the vocabulary chosen by the author of a relevant document. The lack of overlap between query and document terms causes a failure of the document to be retrieved. Query expansion is a technique that addresses this problem by expanding the query with terms related to the information need.

Query expansion is based on the assumption that if two terms tend to co-occur with the same terms, they tend to be related. Relevance feedback [109] is a method by which a query is modified using information derived from documents whose relevance to the query is known. Typically, terms co-occurring with query terms in documents known to be relevant are added to the query. Expansion with related words improves query effectiveness.

Local feedback [9] differs from classic relevance feedback in that no relevance judgments are available. Rather it is assumed that the top retrieved documents are relevant. The technique is carried out in the following way:

- Perform a search on a given query and retrieve a set of documents.

- Assume the top retrieved documents are relevant. Select a set of terms from the top documents via a metric based on frequency of occurrence in those documents.

- Expand the query and perform a second search.

Local Context Analysis (LCA) [134] is similar to local feedback and has been shown to be more effective. It differs from local feedback in that it combines global and local text analysis to identify expansion terms. More specifically:

- Local: Identify concepts in the top-ranked *passages*. This reduces the chance of expanding with words only spuriously associated with query terms due for example, to their appearance at the end of a relevant document.

- Global: Select concepts based on co-occurrence with query terms. Reward them for co-occurring frequently, but penalize them for occurring frequently throughout the corpus.

1.2.3.2 Structured Queries

An additional means by which queries can be made to more accurately reflect their intended meaning is to add structure to them via query operators [30]. The structure that the operators impose makes more explicit the way in which the words in the query are being used, or how important the user feels a word or group of words is to the concept being conveyed. For example, the phrase “greenhouse effect” refers to a concept more specific than is described by either of the words alone. It is more likely that a document containing the phrase would be related to global warming than one containing only the non-adjacent words “greenhouse” and “effect”. The user could make this distinction clear by using a proximity operator which specifies the maximum distance separating the words. For example, #1(greenhouse effect) indicates that the words “greenhouse” and “effect” should be found no more than one word apart. Any occurrences of the words that exceed the maximum separation are given less or no weight.

1.2.3.3 Term Co-occurrence

In addition to the techniques described above, classification methods have been applied to improve IR effectiveness. The *Cluster Hypothesis* [107] states that documents that are closely related will tend to be relevant to the same requests. Many approaches have focused on document clustering for browsing and retrieval. However, our work is more closely related to the application of these techniques to the task of identifying semantic relationships between words.

Words take their meaning from the context in which they are used, thus one might expect that words occurring together are related. The assumption that follows is that analysis of co-occurrence statistics should allow us to infer semantic relationships. Measures of association between words are typically calculated based on the number of documents or text windows in which the words co-occur. If a word is good for a particular task such as describing an information need or partitioning relevant from non-relevant documents, then closely related words should be good for the same purpose.

EMIM [105] is a measure of association that is often applied to index terms. It is calculated via

$$I(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

This is a symmetric measurement. EMIM measures the extent to which the terms co-occur over and above what would be expected if they were independent. Term association has been applied to tasks such as building thesauri and query expansion. In the cross-language environment, we use co-occurrence analysis for disambiguating query translations. This will be described more fully in Section 6.2.2.

1.3 Multilingual Retrieval

Despite growing interest in access to multilingual information, the bulk of IR research has been done with English. The goal of multilingual Retrieval (MIR) is the development of IR systems designed to perform retrieval in a language other than English (e.g.[58, 46, 84]). The focus has been on establishing whether approaches effective for English can be applied to other languages. When this is not the case, focus shifts to developing tools and techniques for addressing language specific problems. There is a whole host of difficulties that arise when one is dealing particularly with non-Western languages. These issues are related to the three main retrieval processes described in Section 1.2.

The first group of issues is related to object representation. To begin with, character encodings vary across and within languages. English is easily represented via the ASCII code and systems have no trouble processing ASCII. However other languages based on the Roman alphabet have additional orthographic characters typically represented by a byte code larger than 128, which may require special treatment for input, indexing, and display. More problematic is the representation of languages containing non-Latin alphabets such as Arabic and Japanese. There are several ways to encode these character sets and all of them must be supported. This is especially important for languages such as Serbo-Croatian that contain different character sets. Although it may be relatively simple to make these adjustments, it is important that care be taken to ensure that the same character encodings and processing are applied to both documents and queries.

Even when character encodings and term processing are consistent across text objects, problems may arise from the way in which a user interacts with the system. If special key combinations are necessary for entering diacritics, people may be less inclined to add accents and the like to their query words. This will cause difficulty if words with the same spelling are treated differently based upon whether or not they contain diacritics. For example, in Spanish te and té have different meanings, the former being a pronoun meaning 'you' and the later the noun 'tea'. If the indexing routine strips all terms of diacritics by default, there will not be a problem since the meanings are conflated. However if diacritics are preserved, matching query and document terms will be more complicated.

Larger problems associated with the processing of non-Western languages are not so easily addressed. The process of stemming (described in Section 1.2.1), is much more complicated for languages such as Hebrew and Arabic which have a complex morphological structure [8, 130]. Inflectional morphology refers to the possible variations on a base form and languages can be classified according to the extent to which they use inflectional morphology. On one end of the scale are languages like Chinese that have almost no inflectional morphology and on the other end of the scale are languages such as Mohawk where the burden of expressing grammatical relationships is placed on word formation [11]. In Hebrew, for example, nouns and verbs can be inflected or conjugated to indicate attributes such as gender, tense, or number. These conjugated forms may also be compounded by the addition of a variety of suffixes. This complexity makes the application of simple suffixing rules for stemming like those applied to Western languages ineffective. In addition, written texts for non-Western languages may be non-vocalized. This means that only consonants are displayed in written texts so that the identity of words must be inferred from clues such as context. This may be relatively easy for a human to accomplish, but is far more complicated to encode in a computer algorithm.

The identification of words is relatively straightforward in many Latin-based languages (i.e. Spanish, English) since words are delineated by spaces. However, the existence of word-boundary delimiters does not guarantee ease of word identification. In agglutinating languages such as German, many words are compounds. Words that would otherwise be treated as separate words or even different parts of speech in languages such as English, are treated as part of one complex word. More complex analysis of agglutinating languages must be done to identify lexical units.

Word identification can be more complicated in the case of languages like Chinese and Japanese that do not explicitly represent word boundaries. Chinese characters are ideograms for which the meaning varies depending upon whether they are used alone or as part of a multi-character concept or phrase. One could ignore collocations and treat each individual character as a word. However, if the goal is to try to identify the combinations of characters that the writer intended, then the employment of a segmenter is necessary. A segmenter analyzes the text by measuring for the current character or combination of characters, the likelihood that the next character should be part of the current word or whether it should be terminated. Detailed discussions of segmenters for Chinese and Japanese can be found in [131, 88].

The identification of word boundaries is especially important in the context of indexing. Retrieval performance is effected by the choices that are made in choosing an indexing language. *Exhaustivity* and *specificity* [105] of an indexing language are measures of the range of topics that the language can describe and the detail with which they can be described. High specificity tends to produce high precision and low recall, while low specificity has the reverse effect. High exhaustivity leads to high recall and low precision, while low exhaustivity yields low recall and high precision. The majority of research on the effect of indexing language on retrieval has been done in English. However, the increased interest in multilingual systems has led to an increase in similar studies for other languages, some of which can be found in [46, 5, 4, 3, 80, 84, 63, 128, 129].

Typically, retrieval systems attempt to balance the trade-off between recall and precision. This can be done a priori by careful construction of the indexing language or by the application of query processing techniques at search time. Character-based indexing (every character is an index term) for a language such as Chinese, for example, can be considered to be highly exhaustive. In this case, post-coordinate indexing can be applied. Post-coordination allows the co-ordination of index terms to be made at search-time via the specification of query term relationships with structured queries as described in Section1.2.3.2. In addition, other approaches such as relevance feedback and query expansion can also be applied. In [47], Fujii and Croft compare query processing strategies for Japanese and English databases. Their results show that techniques for improving retrieval effectiveness for English are also effective for Japanese. However, the strategies for applying the techniques were quite different.

A retrieval system must be tailored for the language upon which it operates. This is due to the many language-specific characteristics such as character representation, word morphology, and language typology. Techniques such as query expansion have been applied successfully to several languages [113, 3, 4, 23], suggesting that some IR methods may be language independent. However, work like that of Fujii and Croft [47] shows that strategies for applying these techniques still vary across languages. Determining which techniques are applicable to languages other than English and development of language specific methods is the role of multilingual IR. The study and development of cross-language IR systems benefits from the lessons learned in MIR studies. However, the focus of cross-language retrieval is on methods for bridging the gap between languages. The primary issues that must be addressed in a CLIR system are the ambiguity associated with translation and the limited availability of dictionaries and other such resources. These problems are discussed in the next section.

1.4 Cross-language Retrieval Problem

Cross-language information retrieval aims to develop tools that in response to a query posed in one language (e.g. Spanish), allow the retrieval of documents written in other languages (e.g. French). There are several approaches one could take to solve this problem. Each of them amounts to generating a translation or more appropriately, an approximate translation of the document into the language of the query or of the query into the language(s) of the documents being searched. Unlike a machine translation system, the goal of a CLIR system is not to generate exact, syntactically correct representations of a text in other languages. It is rather to cull through the tremendous number of electronic texts and to select and rank those documents that are most likely related to a query written in another language.

There are two main hurdles to effective CLIR. The first is reducing the ambiguity associated with mapping the purport of a text object across languages. That is, a CLIR system must address both within language ambiguity and cross-language ambiguity so that the gist of the original text object is preserved. Translation ambiguity is a result of erroneous word translations, failure to translate multi-term concepts as phrases, and the failure to translate out-of-vocabulary words. Previous work ([12] and [68]) showed that ambiguity greatly reduces the effectiveness of cross-language retrieval in comparison to monolingual retrieval with the same queries. The effects and types of ambiguity effecting CLIR are discussed more thoroughly in Chapter 5.

The second hurdle is addressing the limited number of translation resources such as aligned corpora, bilingual dictionaries or word lists, and Machine Translation (MT) systems. Availability of lexical resources varies considerably and there are virtually no resources for some pairs of languages. Availability depends upon several factors including the commercial viability of producing the resources, proprietary rights, and cost. Cost may be measured by the charge for purchase or use of the item, the degree of preprocessing necessary before it can be applied to the task, and the compute time needed to employ the item. The difficulties associated with the application of the lexical resources given above are described in more detail in Chapter 2.

In this work, a dictionary approach is taken, augmented by statistical analysis for ambiguity reduction. The application of statistical dictionary-based techniques is important because they do not depend upon complex lexical databases or special resources such as parallel or comparable corpora. All of the information we need comes from the databases being searched and from the dictionaries. We employ bilingual dictionaries because they are more prevalent and simpler to apply than other lexical resources. An effective approach to CLIR based on machine readable dictionaries (MRD), gives us a starting place for retrieval across languages for which there is no commercial incentive to create complex MT systems. Furthermore, we investigate a *transitive* translation approach, where a third language is employed as an interlingua between the source and target languages. Results suggest that this is a viable means of performing CLIR between languages for which no bilingual dictionary is available. Employing a transitive translation strategy to a dictionary-based approach provides the foundation for a cross-language architecture that can be quickly re-targeted to include new language pairs. The next section describes the contributions of this research and provides a description of the organization of the rest of this dissertation.

1.5 Research Contributions

This research identifies the problems confronting the task of cross-language retrieval and explores a number of techniques for addressing them. The techniques are applied to a cross-language approach based on automatic dictionary translation between Spanish/English and Spanish/French. However, these techniques are general enough to be applied to other languages and to approaches relying on resources other than dictionaries. In addition, the research is carried out using the INQUERY information retrieval system, but the techniques presented in this thesis are general enough to be used in other IR systems.

Developing techniques for CLIR based on machine-readable dictionaries is significant for two reasons. First, there are MRDs for many commercially-important languages. Resources such as aligned corpora are less prevalent. Second, MRDs have fewer of the disadvantages associated with other multilingual resources. MRDs are more readily available, less expensive, and require less work to prepare than other resources. This implies that a CLIR system based on MRDs could be quickly retargeted to new language pairs merely by the addition of a new dictionary.

Preliminary work that indicates that CLIR via transitive MRD translation is a viable approach increases the significance of this work by making retrieval possible between languages for which there are no direct translation resources. In other words, we can perform retrieval between two languages even if there are no available resources that supply word correspondences.

This research makes the following contributions to the field of information retrieval:

- An analysis of causes of ambiguity associated with the task of cross-language retrieval.

- Practical and effective statistical techniques for CLIR which rely as little as possible on scarce resources. The techniques are robust and capable of significantly improving retrieval effectiveness relative to simple word replacement. Application of these techniques brings CLIR effectiveness near that of monolingual retrieval.

- Comparative analysis of different approaches to the task showing, contrary to popular belief, that effectiveness is not dependent upon linguistic analysis. The statistical techniques described in this thesis are as effective as far more complicated commercially available machine translation (MT) systems. This suggests that CLIR systems can be adapted to many languages with very little effort, unlike MT systems which require a significant effort for each new language pair.

- A general, effective approach to CLIR via machine readable bilingual dictionaries.

- The approach provides the foundation of a general architecture for effective CLIR in a general domain.
- *Transitive translation* via MRD increases the significance by circumventing the problem of limited resources.

1.6 Organization of the Dissertation

This dissertation describes an effective approach to cross-language retrieval via machine readable dictionaries augmented by statistical techniques that significantly reduce the effects of translation ambiguity. The approach is robust, cheaper to port than other approaches, and is as good as or better than others that apply more complex techniques and resources such as machine translation or aligned corpora.

Chapter 2 reports on the lexical resources employed in cross-language retrieval and discusses their advantages and disadvantages. Approaches to machine translation are also described. Chapter 3 discusses previous work in the field and other related work. Our experimental methodology and evaluation measures are explained in Chapter 4. Chapter 5 discusses the factors effecting ambiguity and techniques for resolving ambiguity are detailed in Chapter 6. The viability of a transitive translation approach is explored in Chapter 7. Finally, conclusions and directions for future research are summarized in Chapter 8.

CHAPTER 2

LEXICAL RESOURCES AND MACHINE TRANSLATION

An information retrieval system implements a method of representing queries and documents typically written in one language. This representation facilitates comparisons for the evaluation of the relevance of a document to a query. A system capable of bridging the language gap must be able to facilitate comparisons between queries and documents written in different languages. An obvious approach is to translate documents to the language of the query or vice-versa. The former is impractical for two reasons. First, full document translation is time consuming. It can take weeks of compute time to translate only a couple MB of documents. This is more of a problem as the number of documents in a collection grows and takes far too long to be applied in real-time. Second, even if the translations are done off-line so as to be transparent to users, there is the issue of storage space. A system taking this approach would have to know a priori all languages in which a person would like to search and would have to store the translations of its collections into each of those languages. This becomes more unrealistic as the collection to be searched grows, especially in an environment like the web which has access to terabytes of information. Thus, approaches to CLIR have focused on the translation or “pseudo-translation” of queries. We use the term *pseudo-translation* because an approximate translation or gist of the query is sought. Document translation approaches are described in more detail in Section 3.1.3.

Methods for translation have focused on three types of resources: *aligned corpora* for generating a translation model, *machine readable dictionaries*, and *machine translation* (MT) systems. The last 40 years of research in machine translation has led to the development of a number of commercially available MT systems, but these systems are treated as a resource, rather than a goal for three reasons. First, effectiveness of these systems varies considerably, especially when the domain is not limited. Second, building and maintaining an MT system requires considerable effort for each language pair. Finally, the research done for this thesis shows that the degree of complexity employed in MT systems is not necessary for effective CLIR. Sections 2.1,2.2 and 2.3 describe each of the above resources in more detail. Specific approaches to CLIR are discussed in Chapter 3. Section 2.4 gives a summary and discusses the reasons that an approach based on machine-readable dictionaries was chosen for the work in this thesis.

2.1 Machine Readable Dictionaries

Bilingual dictionaries and word lists are the most readily available resource for cross-language retrieval, providing a simple, straight-forward mapping between languages. Dictionary translations are based on these simple mappings. Translation is not used here in the sense of deep linguistic analysis. The terms in one language are merely replaced with the dictionary definition of those terms in another language.

Machine readable dictionaries often require some degree of preprocessing before they can be applied. This is primarily because dictionaries are designed for use by humans. Dictionary mark-up identifies information such as head-words, part-of-speech, and word usage, but may be inconsistent. These inconsistencies are easily differentiated by a person, but they can make computer analysis more challenging.

A dictionary typically does not contain an entry for every word in the language, but rather has entries for the base forms of words. To find the meaning of a particular word, one must locate the entry for the base form of that word. For example, to find the meaning of *retrievability*, the entry for *retrieve* must be located. A typical dictionary entry consists of a head-word (base form) followed by information which can include pronunciation, derivation, inflected forms, meaning, and idiomatic usage, in addition to translation equivalents. In the case of a word that can function in more than one way, a separate entry or sub-entry for each part of speech with its corresponding information will also be given. Tables 2.1 and 2.2 show example entries from the Collins French/English dictionary. The first is an example entry from the machine readable form of the dictionary including typesetting mark-up. The beginning and end of a head-word is identified by ">H<" and ">X<", respectively. Special characters having ASCII codes larger than 128 are given representations that must be substituted (e.g. á would be shown as "a>ac<" and í as "i>ac<"). >LOZ< indicates a difference sense of the word. One final example of typesetting codes is that the general subject field of a particular meaning may be shown in lowercase flanked by ">SR<" and ">R<".

The second table shows the entry for the same word as is found in a printed version of the dictionary. As one can see, the typesetting codes have been replaced with formatting. For example, individual entries are listed for homographs and are distinguished by superior boldface digits.

As is the case for other linguistic resources, ambiguity is a significant problem in applying dictionaries to translations. There may be few clues as to which context should apply for a particular translation. Even if that information is available, it is not trivial to use this knowledge automatically and mark-up can be inconsistent, making it difficult to exploit. While rules for applying knowledge may be effective in some contexts, it is typically not the case for all contexts. Minor changes to resolve errors and inconsistencies often lead to the introduction of other problems. In addition, word-by-word translations may not be appropriate for all concepts, so phrasal entries are a valuable resource for their correct translation. Unfortunately, headword entries of phrases may be scanty if included at all. The Collins English/French MRD employed in this work contains 3850 English phrases and 68 French phrases of which

Table 2.1. Dictionary entry for the word *row* from the Collins machine-readable English/French bilingual dictionary. Typesetting codes are included here, but would not be shown in the printed version.

```
>H<row>X<>sup1<
>[<r>u11<u>]<
>I<n
>R<(>SR<line>R<)
>R<range>ac<e >I<f>R<;
(>SR<of people, seats>R<, >SC<KNITTING>R<) rang >I<m>R<;
(>SR<behind one another: of cars, people>R<) file >I<f>R<
>LOZ< >I<vi
>R<(>SR<in boat>R<)
>R<ramer;
(>SR<as sport>R<) faire
de l'aviron>R<
>LOZ< >I<vt
>R<(>SR<boat>R<)
>R<faire aller a>gr< la rame >GI<or >R<a>gr< l'aviron>R<;
>B<in a > <
>R<(>SR<fig>R<)
>R<d'affile>ac<e>R<.>QL<
>H<row>X<>sup2<
>[<rau>]<
>I<n >R<(>SR<noise>R<) vacarme >I<m>R<;
(>SR<dispute>R<) dispute >I<f>R<, querelle >I<f>R<;
(>SR<scolding>R<) re>ac<primande >I<f>R<, savon >I<m>R<
>LOZ< >I<vi >R<(>SR<also>R<: >B<to have a > <>R<) se disputer,
se quereller.>QL<
```

only 3831 and 52, respectively, were correctly identified by the dictionary parsing software. These are relatively small numbers of phrases when one considers the number of multi-term concepts in a language that are more accurately translated as a unit, rather than word-by-word. Furthermore, the translation of out-of-vocabulary (OOV) terms like specialized terminology and idioms is particularly challenging since they are often not included in dictionaries. When they are included, they are given as examples of word usage and not as independent headwords making them difficult to recognize. Although some specialized dictionaries may contain technical terminology as headwords, they are less prevalent than general dictionaries.

Table 2.2. Entry from the Collins printed English/French bilingual dictionary. Typesetting codes are replaced by formatting.

row$^{\mathbf{1}}$[rəʊ] *n [objects, people]* rang *m*; (*behind one another*) file *f*,ligne *f*;*[houses, trees, figures]* rangée *f*, *[cars]*file; (*Knitting*) rang. **In the front**∼ au premier rang; **sitting in a** ∼ assis en rang; **4 failures in a** ∼ 4 échecs de suite.

row$^{\mathbf{2}}$[rəʊ] **1** *vt boat* faire aller à la rame *or* A l'aviron. **to** ∼ **sb across** faire traverser qn en canot. **2** *vi* ramer. **to** ∼ **away/back** s'éloigner/revenir à la rame; **to go** ∼**ing** (*for pleasure*) faire du canotage; (*Sport*) faire de l'aviron. ◇ **rowboat** *or* ◇ **rowing boat** *n* canot *m* (à rames). ◇ **rower** *n* rameur *m*, -euse *f*.

◇ **rowing 1** *n* canotage *m*; (*Sport*) aviron *m*; **2** *adj*: ∼**ing club** club *m* d'aviron.

◇ **rowlock** [r^{a}lək] *n* dame *f* de nage.

row$^{\mathbf{3}}$ **1**[raʊ] **1** (*noise*) tapage *m*, vacarme *m*; (*quarrel*) querelle *f*, dispute *f*; (*scolding*) réprimande *f*, savon * *m*. **To make a** ∼ faire du tapage; **to have a** ∼ se disputer (*with* avec); **to give sb a** ∼ passer un savon à qn *; **to get (into) a** ∼ se faire passer un savon *; **2** *vi* se disputer (*with* avec).

Word lists contain only lists of head-words and their translation equivalents. Part of speech information may also be included. Because they contain less information than dictionaries, they typically require less pre-processing. Table 2.3 shows an exerpt from an on-line English/Spanish word list from the Activa translation service [1]. This list is simple, listing only headwords and translation equivalents with the format *headword:translation equivalent.* If a word has more than one meaning, a separate entry is listed for each.

Table 2.3. Exerpt from on-line English/Spanish word list listing only headwords and translation equivalents.

row: fila
row: línea
row-column scanning: exploración de líneas
rowboat: bote de remos
rowdy: camorrista
rower: remero
rowlock: sardinel
royal: real
royalist: realista
royalties: derechos de autor
royalty: realeza
RTP (abbr. for: real time protocol): protocolo de emisión

Ideally, a linguistic resource in a general retrieval domain would have two characteristics. First, it would contain vocabulary used in a wide variety of settings. We refer to this as breadth. Second, the resource would contain a representative sample of words from each domain. This is referred to as depth. The coverage of MRDs, while not deep, is broad enough to be used for translations of queries covering a wide variety of topics.

Despite the problems of ambiguity associated with dictionary translation, dictionaries are the most readily available of the resources employed in cross-language retrieval. One can typically find bilingual dictionaries for most commercially-important languages and there are many bilingual dictionaries between English and other less commercially-important languages. Many of these may not exist in machine-readable form, but advances in OCR technology enable the recognition of languages other than English and conversion of text documents to a machine readable form.

The on-line availability of dictionaries [16, 45, 138, 42, 10] is also increasing, including languages such as Chinese, Bulgarian, English, Swahili, Spanish, Turkish, and French. A simple web search found over 70 links to general, multilingual dictionaries on one web page alone [96]. In addition, there is growing interest by many groups to develop projects for which the goal is to make multilingual dictionaries available freely on the web [41, 70, 39]. Although machine readable dictionaries are becoming more widely available, their coverage and quality varies. One would expect this variability to impact effectiveness; however, to what degree has yet to be studied directly. Nonetheless, the greater availability of dictionaries gives dictionary-based approaches to CLIR greater potential for application across more languages.

2.2 Aligned Corpora

Aligned corpora provide a coarser mapping between languages than that given by dictionaries. Aligned corpora are multilingual collections of related documents. They are referred to as aligned because we have explicit knowledge about how documents are related and which documents are related to each other. Finer-grained mappings between document constituents such as sentences or words must typically be inferred. This section discusses two types of aligned corpora for which the type of document association differs: *parallel corpora* and *comparable corpora*.

A parallel corpus is comprised of a set of documents written in one language and the translation of those documents into another language(s). The Hansard Corpus [59] is one example. It contains transcripts of the Canadian parliamentary proceedings written in English and their translations into French. Parallel corpora are primarily the result of large scale translation projects commissioned by large organizations for a particular class of documents. For this reason, few parallel corpora are available and their coverage tends to be domain specific.

Because parallel corpora are generated for human distribution, some degree of preprocessing must be done. For example, the UN corpus [124] was automatically extracted from the UN electronic text archives. This corpus contains English transcripts of UN proceedings from 1988-1993 and translations of those transcripts into French and Spanish. The original archival structure did not give for any document, any indication of which other documents were its translation. Due to the large number of documents (92,000), document alignments were done largely automatically, followed by a manual scan. Alignments were based on conventions for identifying documents such as the custom of assigning a uniform title string to a document and all of its translations. This method was successful for aligning more than 60% of the original text into parallel document sets. However the resulting parallel corpus does contain errors ranging from large discrepancies in the amount of text across documents in a parallel set to complete mismatches. Tables 2.4, 2.5, 2.6 and 2.7 show example parallel documents taken from the English and Spanish portions of the UN corpus. The documents in the first two tables comprise one aligned set. Only a portion of the Spanish document is included for reasons of space. The third and fourth tables contain documents from a second aligned set. These examples are used to illustrate the variability in document alignment quality. There is a large disparity in the length of the first two documents. The lengths of the second pair of documents are nearly the same, differing by about two lines.

Table 2.4. English document of parallel set one from the UN corpus. This document is the mate to the document in Table 2.5.

22 February 1989
ORIGINAL: ENGLISH
GENERAL ASSEMBLY SECURITY COUNCIL
Forty-fourth session Forty-fourth year
Items 31, 72 and 138 of the preliminary list*
THE SITUATION IN KAMPUCHEA
REVIEW OF THE IMPLEMENTATION OF THE DECLARATION ON THE
STRENGTHENING OF INTERNATIONAL SECURITY PEACEFUL SETTLEMENT
OF DISPUTES BETWEEN STATES
Letter dated 21 February 1989 from the Charg d'affaires a.i. of the Permanent Mission
of Democratic Kampuchea to the United Nations addressed to the Secretary-General
I have the honour to transmit herewith, for your information, a press statement
by the three components of the Coalition Government of Democratic Kampuchea (see
annex), approved on 20 February 1989 by HisRoyalHighnessSamdechNorodomSihanouk,
President of Democratic Kampuchea, National Leader of Cambodia and
Head of the Cambodian National Resistance.
I should be grateful if you would have the present letter and its annex
distributed as an official document of the General Assembly, under items 31, 72
and 138 of the preliminary list, and of the Security Council.
(Signed) SISOWATH Sirirath
Charg d'affaires, a.i.

A/44/50.
89-04724 0673e (E)

On the other hand, the documents in a comparable corpus are not translations of each other but are related by topic. The news articles created by the Swiss News Agency (SDA) form a comparable corpus. Switzerland has three primary languages: French, German, and Italian. News stories are written independently, describing the same event, in each of these languages. Table 2.8 shows a German/French document pair from the SDA corpus. Alignment was based on a combination of manual and

Table 2.5. Spanish document of parallel set one from the UN corpus. This document is the mate to the document in Table 2.4. Note: The discrepancy between the document lengths for this set of documents is larger than it appears here. One-hundred thirty-one (131) lines of text of this document are not shown here due to space limitations.

22 de febrero de 1989
ESPAÑOL
ORIGINAL: INGLES
ASAMBLEA GENERAL CONSEJO DE SEGURIDAD
Cuadragésimo cuarto período de sesiones Cuadragésimo cuarto año
Temas 31, 72 y 138 de la lista preliminar*
LA SITUACION EN KAMPUCHEA EXAMEN DE LA APLICACION DE LA
DECLARACION SOBRE EL FORTALECIMIENTO DE LA
SEGURIDAD INTERNACIONAL ARREGLO PACIFICO DE
CONTROVERSIAS ENTRE ESTADOS
Carta de fecha 21 de febrero de 1989 dirigida al Secretario General por el Encargado de Negocios interino de la Misión Permanente de Kampuchea Democrática ante las Naciones Unidas Tengo el honor de transmitir adjunta, para su información, una declaración de prensa de los tres componentes del Gobierno de Coalición de Kampuchea Democrática (véase anexo) aprobada el 20 de febrero de 1989 por Su Alteza Real Samdech Norodom Sihanouk, Presidente de Kampuchea Democrática, Jefe Nacional de Camboya y Jefe de la Resistencia Nacional Camboyana. Le agradecería que hiciera distribuir la presente carta y su anexo como documento oficial de la Asamblea General, en relación con los temas 31, 72 y138dela lista preliminar, y del Consejo de Seguridad.
(Firmado) SISOWATH Sirirath
Encargado de Negocios interino
Anexo Está perfecto. Apruebo totalmente este documento y apoyo cordialmente la muyjustaposición de la delegación de nuestra Resistencia Nacional Camboyana (CGDK) ala JIM II.
Con los más cordiales afectos.
(Firmado) NORODOM SIHANOUK
Presidente de Kampuchea Democrática, Jefe Nacional de Camboya y Jefe de la Resistencia Nacional Camboyana
Beijing, 20 de febrero de 1989

DECLARACION DE PRENSA DE LOS TRES COMPONENTES DEL CGDK
I. En el espíritu de contribuir a la búsqueda de una solución política global al problema de Kampuchea de modo de poner fin a los indescriptibles sufrimientos del pueblo camboyano y con ello asegurar la paz, la seguridad y la estabilidad en el Asia sudoriental, asistimos a la segunda reunión oficiosa de Yakarta (JIM II) en un espíritu de buena voluntad y liberalidad, y no hemos escatimado esfuerzos para que la JIM culmine con éxito.
II. Para contribuir al logro de ese objetivo, la Resistencia Nacional Camboyana -el Gobierno de Coalición de Kampuchea Democrática (CGDK)- ha hecho gala de flexibilidad y ha hecho concesiones desde la celebración de la JIM I, como se indica a continuación:

Table 2.6. English document of parallel set two from the UN corpus. This document is the mate to the document in Table 2.7.

A/47/126 19 March 1992
ENGLISH
ORIGINAL: ENGLISH/FRENCH
Forty-seventh session
Items 69 and 98 of the preliminary list*
REVIEW OF THE IMPLEMENTATION OF THE DECLARATION ON THE
STRENGTHENING OF INTERNATIONAL SECURITY
HUMAN RIGHTS QUESTIONS
Letter dated 19 March 1992 from the Permanent Representative of
Portugal to the United Nations addressed to the Secretary-General
I have the honour to transmit herewith the text, in English and French, of a statement of the European Community and its member States on Mozambique, issued at Lisbon on 17 March 1992 (see annex). I should be grateful if you would have the text of the present letter and its annex circulated as an official document of the General Assembly under the items 69 and 98 of the preliminary list.
(Signed) Fernando REINO
Ambassador of Portugal
Permanent Representative to the United Nations

A/47/50.
92-12996 3036g (E) 200392 /...
ANNEX
Statement on Mozambique made by the European Community
and its member States on 17 March 1992
The Community and its member States, who have been following from the outset the course of the peace negotiations between the Government of Mozambique and RENAMO, welcome the signing of ProtocolIII, concerning the electoral law and citizens rights, which tooks place in Rome, last Thursday, 12March. They hope that this important new development for the process of national reconciliation will be followed by a significant decrease in the intensity of fighting on the ground before the signing of the cease-fire agreement. Recalling their statement of 27 May 1991, the Community and its member States reiterate their support for the efforts of the mediators and encourage the negotiating parties to pursue their efforts towards the bringing about of a final, comprehensive peace agreement.

Table 2.7. Spanish document of parallel set two from the UN corpus. This document is the mate to the document in Table 2.6.

A/47/126
19 de marzo de 1992
ESPAÑOL
ORIGINAL: FRANCES/INGLES
Cuadragésimo séptimo peréodo de sesiones Temas 69 y 98 de la lista preliminar*
EXAMEN DE LA APLICACION DE LA DECLARACION SOBRE EL
FORTALECIMIENTO DE LA SEGURIDAD INTERNACIONAL
CUESTIONES RELATIVAS A LOS DERECHOS HUMANOS
Carta de fecha 19 de marzo de 1992 dirigida al Secretario
General por el Representante Permanente de Portugal ante
las Naciones Unidas
Tengo el honor de transmitir por la presente el texto, en inglés y
francés, de una declaración de la Comunidad Europea y sus Estados miembros
sobre Mozambique, emitida en Lisboa el 17 de marzo de 1992 (véase el anexo).
Le agradeceréa que hiciese distribuir el texto de la presente carta y su
anexo como documentos oficiales de la Asamblea General en relación con los
temas 69 y 98 de la lista preliminar.
(Firmado) Fernando REINO
Embajador de Portugal
Representante Permanente ante
las Naciones Unidas

A/47/50.
92-12999 2022j 200392 200392 /...
ANEXO
Declaración formulada por la Comunidad Europea y sus Estados
miembros sobre Mozambique el 17 de marzo de 1992
La Comunidad y sus Estados miembros, que han observado desde el comienzo el
desarrollo de las negociaciones de paz entre el Gobierno de Mozambique y la
RENAMO, acogen con satisfacción la firma del Protocolo III relativo a la ley
electoral y a los derechos de los ciudadanos, que ha tenido lugar en Roma el
jueves 12 de marzo último.
La Comunidad y sus Estados miembros conféan en que a esa importante novedad
en el proceso de reconciliación nacional seguirá una disminución considerable
de la intensidad de los combates sobre el terreno antes de la firma del acuerdo
sobre cesación del fuego.
Recordando su declaración de fecha 27 de mayo de 1991, la Comunidad y sus
Estados miembros reiteran su apoyo a los esfuerzos de los mediadores y alientan
a las partes negociadoras a proseguir sus actividades encaminadas a la
concertación de un acuerdo de paz completo y definitivo.

Table 2.8. A German/French aligned document pair from the SDA corpus.

USA: Bevölkerungungswachstum 1988 2,3 Millionen Sperrfrist 6 Uhr. Washington, 1. Jan. (sda/Reuter) Die Zahl der Einwohner der USA ist 1988 um etwa 2,3 Millionen auf 246,9 Millionen gestiegen. Wie das Volkszählungsamt am Sonntag in Washington mitteilte, basiert die Schätzung für 1988 auf 3,9 Millionen Geburten, 2,2 Millionen Todesfällen sowie einem Zuwanderungsgewinn von 606 000 Menschen. Seit der Volkszählung von 1980, die 226,5 Millionen Einwohner ergab, sei die Bevölkerung damit um etwa 9 Prozent gewachsen. Für dieses Jahr wird eine Zunahme um 2,2 Millionen Menschen erwartet.
La population américaine a augmenté de 2,4 millions en 1988. Washington, 1er jan (ats/reuter) La population des Etats-Unis était de 246,9 millions d'habitants au 1er janvier 1989, en augmentation de pres de 2,3 millions par rapport a l'annee derniere, a indique le Bureau du recensement dimanche. Les demographes de l'agence federale ont ajoute que la population avait cru de neuf pc par an depuis le recensement de 1980, qui l'avait estimee a 226.545.805 habitants. Le Bureau a calcule ce chiffre en recoupant les 3,9 millions de naissances, les 2,2 millions de morts et une migration nette de 606.000.

automatic techniques. It is assumed that the vocabulary used to describe an event in one language will contain the translations of the vocabulary used to describe the same event in another language. Comparable corpora may be easier to find when one considers the availability of on-line newspaper articles and other types of text from around the world. However, the real difficulty lies in generating document alignments.

The quality of the lexical information and thus the results attained via either of these resources are dependent on the quality of their alignment. Aligning documents is not a trivial task. A parallel corpus has a table that for each document identifies its translation(s). Given the knowledge that document A_s is related to document B_t, it is known that the translations of words in A_s can be found in B_t. However this knowledge alone is not fine grained enough to identify the translations of specific words, so an additional alignment procedure must be employed. The idea behind the aligned corpus approach is that the words used to describe a particular topic or event will be related semantically across languages. When aligned texts are sufficiently large, statistical methods can be applied to infer the most likely translation equivalents.

2.2.1 Aligning Parallel Corpora

Automatic alignment procedures most often use statistical methods. They are typically length-based [52] or cognate-based (string matching) [27]. Statistical approaches can be compute intensive, and those that are relatively efficient make assumptions that often do not hold. For example, length-based algorithms assume that the number of characters in the *source* (the language to translate from) text is roughly the same as the number of characters in the *target* (the language to translate to) text. This is not realistic in many cases and can depend upon how literal the translations in the corpus are, but at the very least the approach can exploit the fact that the ordering of sentences in the target text will be very similar to that of the source text. Gale and Church [52] report their lowest error rates with a length-based algorithm at 4%, but algorithm performance degrades when the languages are very different. Even if the length assumption holds and the sentence-level alignment has a low error rate, it is non-trivial to generate the word correspondences needed for translation. One difficulty is that unlike sentence alignments where text ordering can typically be exploited, word correspondences can occur at any position in a sentence. An evaluation of an approach by Gale and Church [51] showed that 58% of English words from 800 aligned sentences were correctly matched to their corresponding French translations.

An obvious problem with cognate-matching approaches is that they are only applicable to pairs of languages that have words with a close historical derivation. Such an approach could not be applied to more distantly related language pairs like Russian and Spanish.

2.2.2 Aligning Comparable Corpora

Alignment of comparable texts has been based on a combination of manual processing, feature matching, and dictionary mappings [99]. In the case of the SDA corpus, every news article is manually assigned some number of descriptors, each distinguishing some feature of the story. Descriptors identify things such as what country an event took place in, the type of story (e.g. financial) and the subject of the story. Feature matching is based on concepts like the date that an article was written, cognates, and proper nouns. Dictionary matching is based on the use of bilingual dictionaries to generate a mapping between words and their possible translation equivalents. Comparable text alignments have not been effective when based primarily on feature and dictionary mappings, but with considerable improvements when large numbers of manually assigned descriptors are present [18].

2.3 Machine Translation

The third type of resource that has been applied to the cross-language problem is the machine translation (MT) system. The machine translation task is, given text in language A ($text_A$), to generate text in language B ($text_B$) such that the meaning of $text_A$ is preserved in $text_B$. In order to discover the meaning of a text, problems concerning within language word ambiguity, syntactic structure (grammatical as well as ungrammatical), and between-language vocabulary differences must be addressed. Most MT systems employ a combination of lexical and Natural Language Processing (NLP) tools and techniques. Even those relying on statistical approaches, which will be described in Section 2.3.5.3, are being augmented with low level grammatical information to improve poor performance [7]. These include complex dictionaries, word lists, lexical databases, syntactic analysis, morphological analysis, and semantic analysis. The original goal of performing this task entirely automatically has been thwarted by the complexities of natural language. This complexity has resulted in the necessity for some form of user interaction to be provided before (pre-editing), during, or after (post-editing) translation. Sections 2.3.1 and 2.3.2 describe, in more detail, the lexical tools and language analysis techniques employed by MT systems. Specific approaches are described in the sections that follow. More detailed discussions of the resources and techniques employed in MT and descriptions of specific systems can be found in Machine Translation texts such as [69, 93, 92, 7].

2.3.1 Lexical Resources

The dictionaries of an MT system differ from traditional dictionaries as described in Section 2.1 in that they are far more complex. Rather than listing only base forms, they typically list as many word forms as possible. Word lists for languages with complex inflectional morphology can be very long. For this reason, it is often the case that only base forms are listed and then systems apply morphological processing to generate all other forms.

Each word entry must contain grammatical and semantic information, and will often also include syntactic behavior. Grammatical information includes characteristics such as part-of-speech, gender, and number. Issues of semantics can include domain, category such as whether a word represents an action or physical process, and co-occurrence properties may also be included. Semantic properties can be used to guide selection by specifying the contexts in which a word can can be used. Syntactic rules describing how to combine words to generate well-formed sentences, typically cover three kinds of relationships. The first describes the order in which word types should appear. For example, in English, modifying adjectives typically precede nouns. The second type describes what kinds of words can be combined to create certain constructs such as phrases. The third describes the dependency of certain word forms as determined by the other words occurring around them. For example in the sentence "The children are playing", the plural word *children* requires the verb form *are*. In addition to within-language characteristics, there must be information pertaining to word correspondences across languages. Typically this involves monolingual lexicons for analysis or generation and bilingual lexicons that indicate under which circumstances a particular word-pair are considered equivalent.

The translation quality of an MT system depends largely on the quality of its dictionaries. Translations improve with broader coverage of the vocabulary and with the depth to which syntactic and semantic features are described. Given the critical role of dictionaries, compiling and maintaining them for an MT system is a mammoth undertaking. Section 2.3.6 describes this process in more detail.

2.3.2 Linguistic Analysis

Dictionaries alone do not provide enough information for an operational MT system. Because text tends to be ambiguous, some form of analysis must typically be done to identify its meaning. Text is composed of sentences which in turn are composed of phrases and words. Analytic techniques exploit this compositionality of text by following the rules of grammar in an attempt to resolve the following: identify the correct structure of the source language sentence, select the appropriate translation equivalents with which to replace the source language words, and establish guidelines and procedures for converting the source language text into correct target language sentences. The difficulty in analysis is not the application of grammar rules, but in resolving ambiguities that arise when two or more rules can be applied to the same context. Three types of analysis are typically employed to try to address these issues.

The first is morphological analysis which uses the structure of a word to infer certain characteristics. The two types of morphology are inflectional and derivational.

Inflectional morphology describes the ways in which the base form of words can be modified to alter their form yet maintain the same function. An example is applying the *-s* suffix to create a plural as in *card* and *cards.* Derivational morphology describes the modifications that transform one base form into another. For example, the verb *retrieve* can be transformed into a noun with the endings *-ability*, *-er*, and *-al* or to an adjective with the ending *-able.* A typical application of morphological analysis is to identify part-of-speech. In the case of languages like Chinese in which word boundaries are not explicitly represented, character-based morphology can be employed to infer where those boundaries should occur. Another approach is to use word lists in the way of the commercial MT system SYSTRAN [136]. In this case, a list of around 600,000 words is used to find all possible word matches for a sentence. Overlapping matches are then resolved via special algorithms. Compounding, or the combining of words to create new base forms, is also a morphological issue. Identification and decomposition of compounds is especially challenging when working with "agglutinating" languages like German, where the rules for compound formation are very loose. In English we are used to constructing compounds from nouns such as *air traffic controller* or *address book.* This can be complicated by the fact that some compounds are connected by hyphens such as *baby-sitter* or *car-ferry* and that these conventions are not consistently followed.

Syntactic analysis identifies characteristics of the text based on the syntactic rules of the language it is written in. It is concerned with sentence structure and order, and with the relationships between the words. Relationships of interest include the identification of phrasal constructs. The difficulty with syntactic analysis is that there is often more than one way to generate a particular construct. For example, a noun phrase is a group of words that function as a noun and there are a number of ways to form one, including the following: 1) pronoun, 2) determiner noun prepositional phrase, 3) determiner adjective noun, 4) determiner adjective noun prepositional phrase. Other sentential constructs behave similarly, so analysis must proceed by trial and error until a consistent overall structure is produced. *He hit the man on the street with the car.* is a syntactically ambiguous sentence. It could mean *The man on the street was hit with the car.*, *On the street with the car, he hit the man.*, or *Of all the men on the street, he hit the one with the car.*. While other sentences are not ambiguous to a human reader, they require more than syntactic information to analyze correctly. Semantic analysis is applied to resolve such ambiguities.

Semantic analysis is employed to resolve lexical and structural ambiguities and is the most difficult type of analysis to perform correctly. It involves exploiting information such as the domain, category, or sub-category to which words belong to infer the context in which ambiguous words or constructs are being used. Consider the following sentence: *We bought the land with a duck pond.* This is easily understood by a human, but in order to correctly parse this sentence automatically, the knowledge that a *duck pond* is not a form of currency must be encoded in the system. Resolving ambiguities involving polysemous words is even more challenging. In addition, applying semantic constraints can be complicated by several factors. First, unknown words can't be used to disambiguate context. This is true even when the system is limited to a sublanguage. Second, idiomatic expressions don't follow semantic constraints. Third, it is difficult to encode selection criteria in such a way that they can be applied in all cases without violating some constraint(s).

Automatic text analysis is often performed by a parser. The input to a parser is a grammar, a lexicon, and a text. The output is an analysis of the text's structure. Analysis can be done in either a top-down or a bottom-up fashion. Parsers differ by the type(s) of analysis they perform and the depth to which they perform it. However the more complicated and/or recursive the grammar rules are, the more likely there will be multiple ways to analyze the text, of which more than one may be correct. There is currently no single accepted theory of grammar, so the choice of which rules to implement varies from system to system. As described above, the parser may also use semantic or other linguistic information, user input, contextual cues, or real-world knowledge. The latter two types of information are particularly difficult to exploit automatically. The difficulty in using contextual cues is identifying beforehand which ones will be useful later on, since storing all of them is not practical. The task of encoding and effectively exploiting all potentially applicable real-world knowledge is not currently feasible. Assuming that a sufficient amount of information for a particular domain has been encoded relatively effectively, well formed input does not guarantee that syntactic and semantic analysis will succeed at resolving ambiguities. It becomes even less likely when the input is not grammatically correct.

2.3.3 Human Interaction

Despite the tremendous amount of energy that goes into developing lexical resources and linguistic analysis tools, most MT systems rely on some form of human interaction. Depending upon the type of input text, an MT system's output can vary from unreadable to a typical goal of about 80% accuracy[72]. To get a sense of the types of errors that can be made by an MT system, Table 2.9 shows output of machine translations to English of the parallel Spanish document from Table 2.5, by the AltaVista/SYSTRAN [6] and Lernout&Hauspie/Globalink [85] on-line translation services. The original English document is given in Table 2.4. These translations contain errors, but still allow a reader to get the gist of the original document. This may be acceptable for some applications. The type and degree of human interaction tends to vary depending upon the specific goals or needs of the entity that will be using the system output. The human interaction may be requested during translation, prior to submitting input to the system (pre-editing), or after the system has generated its output (post-editing).

Table 2.9. On-line machine translations to English of the Spanish document from parallel document set one (Table 2.5). The translation at the top was done by AltaVista/SYSTRAN and the one on the bottom by Globalink/A&H.

GENERAL ASSEMBLY SECURITY COUNCIL Cuadragsimo fourth period of Cuadragsimo sessions fourth year Subjects 31, 72 and 138 of the list preliminary the SITUATION IN KAMPUCHEA EXAMINATION OF the APPLICATION OF the DECLARATION ON the FORTIFICATION OF the SECURITY the INTERNATIONAL ADJUSTMENT PACIFICO OF CONTROVERSIES BETWEEN STATES Letter of date 21 of February of 1989 directed to the Secretary General by the temporary One in charge of Businesses of the Permanent Mission of Democratic Kampuchea before the United Nations I have the honor to transmit associate, for its information, a declaration of press of the three components of the Government of Coalition of Democratic Kampuchea (it see annexed) approved the 20 of February of 1989 by Its Height Real Samdech Norodom Sihanouk National head of Cambodia and Jefe of the Cambodian National Resistance. He would be thankful to him that he made distribute to the present letter and its annex like official document of the General Assembly, in relation to subjects 31, 72 and 138 of the preliminary list, and of the Security Council. (Signed) SISOWATH temporary Sirirath In charge of Businesses Annexed He is perfect. I approve totally this document and support sincerely the very right position of the delegation of our Resistencia Cambodian Nacional (CGDK) to JIM II. With the most warm affection. (Signed) NORODOM SIHANOUK President of Democratic Kampuchea, National Head of Cambodia and Jefe of the National Resistance Cambodian Beijing, 20 of February of 1989 DECLARATION OF PRESS OF the THREE COMPONENTS Of CGDK ...
ASSEMBLY GENERAL ADVICE OF SECURITY Fortieth fourth period of sessions Fortieth fourth year Fear 31, 72 and 138 of the preliminary list THE SITUATION IN KAMPUCHEA EXAM OF THE APPLICATION OF THE DECLARATION ON THE INVIGORATION OF THE SECURITY INTERNATIONAL ARRANGEMENT PACIFIES OF CONTROVERSIES AMONG STATES Date letter February 21 1989 directed to the General Secretary for the one Taken charge of interim Business of the Permanent Mission of Democratic Kampuchea before the United Nations I have the honor of transmitting enclosed, for their information, a declaration of it presses of the three components of the Government of Coalition of Democratic Kampuchea (see you annex) approved February 20 1989 for Their Real Highness Samdech Norodom Sihanouk, President from Democratic Kampuchea, National Boss from Cambodia and Boss of the Cambodian National Resistance. He/she would thank him him to make distribute the present letter and their annex as I document official of the General Assembly, in connection with the topics 31, 72 preliminary clever y 138 de la, and of the Council of Security. (Signed) SISOWATH Sirirath In charge of interim Business Annex It is perfect. I approve this document and support totally cordially the muy justa posición of the delegation of our Cambodian National Resistance (CGDK) a la JIM II. ...

During translation, intervention by a human translator is sought to resolve ambiguities. Systems that request manual input during the translation process are typically developed as part of a translator's workstation. The idea is to reduce the load on a human translator by performing non-ambiguous translations automatically and requesting user input when ambiguities cannot be resolved. This type of interaction is especially useful when there are words in the text that cannot be found in any of the dictionaries.

Pre-editing can merely consist of adding special mark-up to identify constructs such as proper nouns or phrases. Alternatively, pre-editing can mean limiting the input text to a controlled vocabulary or sub-language. Many MT systems are limited to a particular domain of discourse via pre-editing because limiting processing to that of a sub-language reduces the level of ambiguity that must be addressed. If the level of ambiguity can be reduced, the system can rely less heavily on automatic procedures for linguistic analysis. Such analyses greatly effect system development and computation costs. Nagao [92] reports that General Motors in Canada hires pre-editing specialists who are trained to write GM's technical documents according to specific rules (p. 132). These rules include limiting vocabulary and sentence structure to yield constructions with unique translations so that MT is more effective. Although pre-editing can be time consuming and may require specially trained staff, it can be cost effective in an environment such as that of GM where technical documents are automatically translated into several different languages.

Although the use of pre-editing is common, according to Hutchins (1992) most operational systems rely on post-editing "to produce acceptable translations" (p. 9). This relies upon a translator to revise system output by removing errors prior to distribution. Revisions are easier to perform in restricted domains for which the translator or the person(s) to whom the translation will be distributed is an expert in the field. This is because there is typically enough context for an expert to get the gist of what the original document was trying to convey. However, post-editing is not necessarily any more efficient than pre-editing. It is possible for the output of an MT system to be so confusing that even a human translator cannot distinguish the intended meaning. This becomes more likely as the level of restrictions on the input text is reduced and especially when MT systems are used in a general domain.

2.3.4 Basic Approaches to MT

MT can be categorized into three main approaches, *direct*, *transfer*, and *interlingual*, that rely primarily on linguistic methods to resolve ambiguity. This section discusses each of them. Variants of these which employ non-linguistic approaches to ambiguity reduction are described in Section 2.3.5.

2.3.4.1 Direct Systems

In a direct MT approach, source language text is translated to a target text representation directly essentially via a word-by-word approach. The target words are typically re-ordered to enforce simple syntactic rules such as "adjectives precede nouns". However, exceptions to these rules must be encoded explicitly in the dictionary.

Only as much analysis of the source text is performed as is needed. In systems taking the direct approach, default translations are used for most words and then repairs to incorrect translations are made when possible. Results generally contain many mistranslations. The direct approach was that applied by most of the early MT systems and variants of some, such as SYSTRAN [119], are still in use today.

Direct systems are designed specifically to translate in one direction between one pair of languages. This is due primarily to the fact that there is little or no separation between the declarative and procedural information that is needed for analysis of the source text and synthesis of the target text. This dependency upon target language knowledge for analysis of source text and the converse for the generation of target text, requires significant effort either to perform translation in the reverse direction or to add a new language to the system.

2.3.4.2 Interlingua

Interlingual systems employ an *interlingua* or intermediate representation that attempts to convey meaning in a language-independent fashion. First, the source language text is analyzed and converted to the interlingual representation. Target text is produced from this representation via the application of target language generation procedure(s). Strictly speaking, the analysis and generation components are completely independent so that any analysis module can be linked to any generation module, thus facilitating the development of multilingual systems. Theoretically, the advantage of the interlingual system is that to add a new language, only one new analysis module and only one new generation module need be added. However, in reality it is not trivial to design an interlingual system for two reasons.

The first difficulty lies in the degree of complexity in developing an interlingua even for closely related languages. The interlingua must include all of the characteristics of all of the languages it will represent, even those characteristics that are not shared by all of them. English, for example, shares neither the gender expressions of Spanish nor the conveyance of social status or respect of Japanese. Defining a universal interlingua has been an open problem studied for several centuries. John Wilkins, the seventeenth century English mathematician, was one of the first to attempt this task [118].

Second, it is difficult to design analysis and synthesis modules. The difficulty with designing synthesis modules is in deciding what information to extract from text in order to build the most language-neutral representation, and how best to extract it. Synthesis modules are challenging because all of the source language nuances which can greatly effect target language choices have been stripped away from the interlingual representation. In fact, most operational systems do not have interlinguas capable of representing all of the conceptual distinctions of languages. Rather, the systems rely on the application of contextual information or real-world information to resolve translation ambiguities. The complexity of encoding this type of information typically restricts the use of interlingual systems to restricted domains. The addition of real-world information is a step towards the development of knowledge-based systems which are discussed further in Section 2.3.5.

2.3.4.3 Transfer Systems

The transfer approach is a three stage process that typically involves intermediate, syntactic representations for both the source and target languages. The process begins by converting the source text to an intermediate representation for which all source language ambiguities have been resolved. The second stage is to convert the intermediate source representation to an equivalent target language representation. The final stage generates target language text from the target syntactic representation created in the second stage.

As in the interlingual approach, analysis and generation programs for a specific language are independent of other languages. However, because there is no language-independent representation, there is a separate source representation and target representation for each language in the system. Moreover the transfer module, responsible for the conversion between source and target representations, is necessarily a language-dependent, bilingual module. This means that to add a new language, in addition to two new modules (one for analysis and one for generation), two transfer modules must be added for every other language in the system. For example, to add French to an existing Spanish and English system, six new modules must be added: one French analysis module, one French generation module, one French-English transfer module, one French-Spanish transfer module, one Spanish-French transfer module, and one English-French transfer module.

SYSTRAN straddles the line between direct and transfer systems. It relies heavily on complex dictionaries, yet uses a representation based on the parse of a sentence to reduce ambiguities. It is described in more detail in Section 2.3.6.

2.3.5 Non-linguistic Approaches to MT

Non-linguistic approaches are based on the notion that there is some language-independent conceptual framework that humans build, that is based on their knowledge of the world. These approaches attempt to map what is written to phenomena such as entities, events, or actions in the real world.

2.3.5.1 Knowledge-based MT

The first of these approaches is known as the knowledge-based approach and is a variant of the interlingua approach. It attempts to create language-independent representations of the text based on semantic analysis. These representations are augmented by an inference mechanism that applies information contained in knowledge bases and a domain model. Difficulties that must be addressed include deciding which concepts to represent, which relationships between concepts should be represented, how to represent them, and the granularity of the relationships. The complexity of constructing knowledge bases and domain models typically limits systems of this type to restricted domains and sublanguages.

2.3.5.2 Example-based MT

Example-Based methods apply information about the correspondences between source and target language constructs that is extracted from large parallel corpora. The approach is based on the idea that in the context of similar sentences, the translations of many sentences will be simple modifications of earlier translations.

Translations are generated by attempting to match the source language text or text fragments to previously translated examples. This is not an easy task. First, it is difficult to determine the best way to match sentences or sentence fragments. Second, assuming a good match has been made, it is difficult to choose the correct correspondence between constituents. Third, even if the previous two issues are resolved, many good, yet overlapping matches will typically be found, making it difficult to determine which combination of matches produces the best translation. Much of the work in this area has focused on developing matching and selection techniques.

The example-based approach also requires that the parallel corpus be aligned at a much finer grain than that of document level alignment. As discussed in Section 2.2.1, alignment is a non-trivial task. In addition, these systems still need analysis and generation modules to analyze input and to identify dependency relationships in the examples. Moreover, computational efficiency is of particular concern, especially as the size of the example database increases.

2.3.5.3 Statistical MT

Like the example-based approach, statistical MT relies upon the availability of large amounts of parallel text. It assumes that correspondences between source and target language constructs can be identified via statistics gathered from the analysis of the source language texts and their target language translations. Systems employing this approach apply little, if any, explicit linguistic knowledge. They rely instead on features such as word co-occurrence to infer the most likely translations.

One such approach taken at IBM by Brown, et al [21] is to sentence align the parallel corpus and then to calculate two types of probabilities. The first are the probabilities that a given source language word corresponds to a group of target language words. The second, referred to as fertility and collected by matching bigrams, is the probability that the source language word maps to zero, one, or two correct target language words. These probabilities are used to build language models of target text strings that are used as selection criteria for candidate translations. The advantage to this approach is that little time needs to be spent on encoding linguistic information. The disadvantages are those associated with the use of parallel corpora. First, a large amount of text is needed so that the probabilistic models will not be under-specified, yet parallel corpora are not plentiful. Second, parallel corpora are typically domain specific, making it likely that important correspondences between some source and target words will not be represented. Third, parallel corpora are not error-free, so pre-processing is necessary. In addition, although this approach does not rely on the building and maintenance of large, complex dictionaries, Arnold et al. report that the IBM group has begun to add linguistic data to the system in order to improve effectiveness [7]. Additions include morphological information and minor input transformations, after which the rate of correct translation for 100 test sentences rose from 39% to 60%.

2.3.6 An MT System Example: SYSTRAN

The above sections discussed the main approaches to MT as well as some variants. In order to get a better sense of the complexity associated with building an MT system, this section focuses on some of the details of the SYSTRAN MT system. The reason for choosing SYSTRAN is two-fold. First, SYSTRAN is probably the best-known example of a direct system. Over the years, its design has become more modular and has moved in the direction of a transfer-based approach (there is still some dispute about whether it can truly be considered a transfer system [53]). Second, Chapter 6 describes a comparative evaluation of MT systems including SYSTRAN and our dictionary approach to cross-language retrieval.

SYSTRAN employs a sentence-by-sentence approach beginning with word analysis which applies dictionary data to try to resolve ambiguities. This is followed by further ambiguity reduction based on a parse of the sentence. Finally, the parsed sentence is translated into the target language. SYSTRAN relies primarily on large, complex bilingual dictionaries [53, 120] containing word correspondences as well as grammatical and semantic information. These dictionaries are of three types. The first type is a *stem dictionary*, which lists every source language word in its root form and gives all semantic, morphological, and syntactic information associated with each. This information includes for example, grammatical category and semantic descriptors describing the kind of object a word represents or the domain to which it belongs. For some languages, all word forms are listed in addition to root forms. Source language information for each word is complemented by transfer and target information for several target languages. This includes listing translation exceptions to regular transfer rules, form, inflection pattern, article usage information, syntactic information, and lexical routines that are designed to resolve ambiguities for polysemous words.

The second type of dictionary is the *expression dictionary*, which contains information about multi-term expressions as well as special word-specific information. Multi-word expressions such as idioms and collocations, are fused together and then entered into the stem dictionary with the appropriate target translation. Links from individual terms in the expressions to entries in the stem dictionary trigger the execution of rules that enable generation of all inflected forms and alternate spellings. Conditional expressions describing under which conditions to assign a particular target language translation are also stored here. Word-specific information includes parsing expressions and rules for the resolution of ambiguities due to polysemy, multiple syntactic usage patterns, or multiple parts-of-speech.

The third type of dictionary are *specialized dictionaries*. They can be edited by the user and are created to contain technical terminology or words and expressions that are not found in the main dictionaries.

The translation process is carried out in stages. The first stage begins with dictionary look-up for idioms and individual words. If necessary, morphological analysis is performed and compound nouns are identified using contextual information in the dictionaries.

Analysis is carried out in the second stage. The parser uses a number of procedural modules that attempt to resolve homographs, segment sentences into phrases and clauses, identify primary syntactic relationships, and finally, to determine what the subject and predicate are. The parser is deterministic, so even ill-formed input will be parsed. Questionable parses are flagged and then a special filter program is applied to identify parse errors.

Stage three invokes target language translation modules. At this stage, idioms and prepositions not recognizable in the first stage are translated. In addition, since the translation of one word may impact the translation of one or more other words in a sentence, lexical routines are triggered to ensure that the appropriate alternative translations of surrounding words are chosen.

The final stage of the translation process is to generate the correct output text. This includes defining the default translation of words not translated in an earlier stage, production of the correct morphological structure, invoking algorithms that create specialized target language constructs, and rearrangement of target word order.

Performance improvements have been based primarily on modifications to the large dictionaries employed. The difficulty here is that dictionary construction becomes quite complex. For example, Gerber [53] reports that SYSTRAN has more than 2.4 million dictionary entries for twelve languages. SYSTRAN maintains a staff of lexicographers and specially trained bilingual personnel whose job it is to develop and maintain these dictionaries. Compiling the entries is only the beginning and must be followed with entry-by-entry validation and refinement. Even if lexicographers are generating entries as efficiently as possible (Gerber indicates 5000 entries/month is not unheard of), the dictionaries for one language could take 40 person-months to compile. This does not take into account the time needed to validate or maintain them. Moreover, changes introduced to address one problem may result in other unimagined and detrimental effects.

2.4 Summary and Discussion

The cross-language approach described in this thesis is based upon machine readable dictionaries. Dictionaries are preferable to parallel or comparable corpora for several reasons. First, they are more prevalent and can potentially be used cover a broader range of languages than parallel corpora. Second, they may be more applicable in general contexts since they provide coverage of a broader range of concepts than parallel corpora, which tend to be domain specific. Finally, they are easier to apply because they make explicit the correspondences between source and target language words. It is non-trivial to generate alignments for aligned corpora.

MT systems typically rely on various types of bilingual dictionaries to identify word correspondences. While dictionary lookup procedures may be cost efficient, building and maintaining them for an MT system is not. This is due to the fact that MT effectiveness improves with the complexity of dictionary entries. However, the effectiveness of a particular dictionary structure for one system does not guarantee that it will be effective for another system. For example, the Taum Meteo system [121] is a very effective transfer system limited to meteorological texts. Development of the Taum Aviation system [73] was based upon Taum Meteo. It was far less effective than Taum Meteo and the project was canceled when it became too expensive to continue to build its dictionaries. In addition, the type of language analysis performed by an MT system greatly effects development and computation costs. There is no generally accepted approach to syntactic analysis, so determining the best methods for a given system and language is time consuming. Performing semantic analysis automatically is especially challenging for two reasons. First, it requires the development of a representation that encodes the essential meaning of text. Second, a set of rules capable of extracting this information and for distinguishing context must be designed. Neither of these tasks is trivial. The deeper the language analysis the greater the computational complexity.

The differences between the three main MT approaches lie in the complexity of the analysis, transfer, and generation modules. The direct approach is on one end of the spectrum, requiring little if any analysis or generation beyond replacement with target equivalents. The interlingual approach requires considerable analysis in order to achieve a language independent representation. Transfer approaches fall somewhere in the middle, requiring less complexity in the transfer stage as the degree of text analysis increases in the preliminary stage. Existing MT systems vary in their effectiveness with even the most effective ones yielding far from perfect translations. Developing an MT system is time consuming and each new language pair requires a significant new effort.

The reason for preferring a cross-language retrieval approach based on MRDs to one relying on MT may not at first be obvious. It is better understood when one considers the considerable difference in their goals. Unlike a machine translation system, the goal of a CLIR system is not to generate exact, syntactically correct representations of a text in other languages. It is rather to cull through the tremendous number of electronic texts and to select and rank those documents that are most likely related to a query written in another language. An approximate translation that captures the gist of the original query is typically sufficient for the latter task, while the former task requires considerably more effort to achieve.

This is not to suggest that an existing MT system that is effective for the language-pair of interest should not be applied to the cross-language problem. However, there are several factors that make a dictionary-based approach to CLIR more attractive.

First, considerably more effort is required to develop new MT systems or to add a new language pair to an existing system than is necessary to apply MRDs. In addition, there are considerably more bilingual dictionaries than there are language pairs covered by MT systems. Furthermore, Chapters 6 and 7 give supporting evidence to show that the level of linguistic analysis performed by MT systems is not necessary for effective CLIR. It also suggests that a transitive translation approach is a viable means of addressing the lack of translation resources. These issues suggest a dictionary-based cross-language approach that is quickly retargetable and has greater potential for application across more languages.

Our dictionary approach to CLIR will be discussed in detail in Chapters 6 and 7. However, in the next chapter we first present a discussion of related work and approaches that others have taken to the CLIR task.

CHAPTER 3
RELATED RESEARCH

Recall from the previous chapter that methods for CLIR have focused primarily on pseudo-translation of queries via employment of machine readable dictionaries and thesauri, or aligned corpora and comparable document collections to generate a translation model. Others have utilized machine translation systems to attempt a strict translation of either queries or documents. This chapter discusses these types of approaches as well as others.

3.1 Generating Translations

Although cross-language retrieval research only began to gain momentum in the last couple of years, Salton [109] discussed it as early as 1970. This early approach, described in Section 3.1.1, was based on *manual indexing*. Recent approaches investigate the use of automatic methods relying primarily on *machine translation*, *parallel corpora or comparable corpora*, and *bilingual dictionaries* to generate approximate translations, but others have taken simpler approaches such as cognate matching. Although each of the approaches has shown promise, they are affected to different degrees by problems associated with the limited availability of resources and translation ambiguity. The following sections discuss each of these approaches in more detail.

3.1.1 Manual Indexing

Salton's approach was to manually assign thesaurus classes to the terms contained in a small collection of French documents and their translations. Groups of related words from each language were placed in individual classes in such a way that corresponding groups from both languages were assigned the same class identifier. The thesaurus classes acted as an interlingua between the two sets of documents. Term weighting was also used, with terms judged to be better discriminators assigned higher weights. Preliminary studies seemed promising, but CLIR was less effective for French queries and English documents: the variability of French (in comparison to English) caused the translations of English terms from a single class to be mapped to French terms spanning several other classes. Some terms had no correct mapping from French to English, and vice-versa. For example, assume English term $word_{e1}$ from $class_{10}$ has French translations $word_{f1}$ and $word_{f2}$ from $class_{10}$ and $class_{23}$, respectively. A $word_{f2}$ in a French query would be replaced with English terms from $class_{23}$, missing the correct translation in $class_{10}$. This resulted in there being no means of expressing some concepts in one of the languages. Results from this study were encouraging, showing greater than 90% of monolingual retrieval effectiveness. Although the test collection was very small by current standards, the main drawback to this approach is that it is unrealistic to manually index large databases.

3.1.2 Cognate Matching

Buckley, et al. [23] took a much simpler approach to the cross-language problem. Noting that many words in French and English share many cognates, English query terms were treated as potentially misspelled French words. No dictionary was used, rather a simple string matching algorithm was applied. First, a list was constructed of the French words occurring five times or more in the corpus to be searched. To match English words to French words on the list, the algorithm modified the English by adding or deleting one letter, or one of two equivalence classes of letters. For example, one of the equivalence classes was defined to cover k sounds such that any combination of c, k, or qu could substitute for any other. When a match was found, the matching French word(s) was(were) added to the query. This simple approach worked moderately well achieving 60% of monolingual performance. This is in the range of expected effectiveness via simple machine readable dictionary translation. An additional experiment was run in which query expansion was applied after running the matching algorithm; however, expansion gave no significant improvement in effectiveness.

The approach employed here suggests that low level retrieval effectiveness can be achieved with very little lexical knowledge. However, it can only be applied to languages with similar etymologies and that have remained essentially the same in form. In addition, it remains to be seen whether statistical or other techniques can be used to reduce the effects of ambiguity and bring effectiveness to an acceptable level.

3.1.3 Machine Translation

Machine translation (MT) systems have been employed to translate documents [95], but the approach is impractical for most environments. Oard states that it took one month of compute time to translate roughly 250MB of German documents into English using five SPARC stations. This inefficiency is one reason why researchers tend to focus on translating queries. Query translation via MT is a viable alternative [50, 103], but MT systems tend to need more context than is in a query for accurate translation. MT effectiveness also relies on syntactic analysis and queries are often not syntactically correct. Another disadvantage of the approach, as discussed in Section 2.3.4, is that development of a system requires an enormous amount of time and resources. Even if a system works well for one pair of languages, each new language pair requires a significant new effort.

3.1.4 Parallel and Comparable Corpora

Parallel corpora are being used by several groups e.g.[81, 35, 26]. Landauer and Littman proposed a method for cross-language retrieval via Latent Semantic Indexing (LSI) [49]. LSI is based on the *vector space* retrieval model in which documents are represented as vectors of terms. It differs from the vector space model in that a reduction on the term-by-document matrix is performed via singular value decomposition. The underlying assumption is that the dimensions of the resulting matrix are representative of "core" or "basic" concepts of discourse. In Landauer and Littman's work, a small parallel corpus of English documents and their French translations were concatenated to create a collection of bilingual documents. LSI was then applied to create a bilingual indexing space for which it was assumed, the resulting dimensions represented "core" concepts that are language independent. Their method has been successful at retrieving a query's translation in the top 10 documents. However, the collection used was small, containing 2482 paragraph-length documents from Canadian Parliamentary proceedings and recent work has not shown it to be effective on the traditional retrieval task.

The work of Carbonell [26], is based on the *generalized vector space model* (GVSM). GVSM differs from the vector space model in that it uses documents as the basis to represent terms. Query-document similarity is evaluated via

$$sim(\vec{q}, \vec{d}) = cos(A_t\vec{q}, A_t\vec{d}),$$

where $\vec{q}$ and $\vec{d}$ are query and document vectors and A is the term-document matrix. The extension of GVSM to the cross-language environment relies on the use of a parallel corpus. Two matrices, A and B, are created, where A is a term-document matrix for the source language and B is a term-document matrix for the target language. The corresponding columns of A and B represent the matching document pairs in the parallel corpus. Query vectors in the source language are compared to documents in the target language via the following similarity metric:

$$sim(\vec{q}, \vec{d}) = cos(A_t\vec{q}, B_t\vec{d}).$$

The system has been tested with some success on a corpus consisting of about 2000 paragraphs from the UN parallel corpus [56]. However, the evaluation is somewhat unrealistic. First, the queries were hand constructed and fine tuned for the document collection. Second, the document domain was narrowed further from general UN activities to only documents containing at least ten occurrences of the term "Unicef". The test collection was then generated by subdividing those documents into paragraph-length text fragments. In addition, like LSI, it has not been shown to be effective on larger, more realistic collections.

Another method that relies on parallel corpora has been suggested by Dunning and Davis [44]. Text objects such as documents and queries are represented as feature vectors. The technique involves the linear transformation of the representation of a query in one language to its corresponding representation in another language via a translation matrix, T. Given the feature vectors for a set of documents and their translations, T is derived by solving large systems of linear equations. The transformation is similar to the singular value decomposition calculated by LSI. However, tests of the effectiveness of the method have been limited by its computational complexity.

Davis and Dunning also developed several other approaches to query translation for cross-language retrieval, all relying on a parallel corpus. In two methods [37], "translations" are performed by replacing English query terms with high frequency or statistically significant terms from the Spanish side of the parallel corpus. First, one hundred English documents were retrieved in response to an English query. Spanish query term pseudo-translations were then extracted from the Spanish documents that were parallels to the retrieved English documents.

A third technique uses Evolutionary Programming (EP) [36] to optimize queries generated by one of the first two methods or via word-by-word machine readable dictionary (MRD) translation. It first modifies queries by randomly adding or deleting query terms. Optimization is done by evaluating query fitness after each round of mutations, and selecting the "fittest" to continue to the next generation. The fitness judgment was based on a comparison of retrieval results from a parallel corpus. An ititial retrieval was performed for the English queries on a sentence alignment of the UN parallel corpus. Spanish queries were generated by one of the methods described above, then retrieval was performed on the Spanish side of the aligned corpus. The most fit Spanish queries were considered to be those that retrieved the most Spanish sentences that were parallel to the sentences retrieved in response to the original English query. The evolutionary programming approach was the most effective of the above methods, but results were disappointing, with each performing well below the word-by-word translation baseline.

Davis later used part-of-speech tagging to select the best Spanish translations for English query terms from a machine readable dictionary [35]. A parallel corpus was then used to further disambiguate the translated queries by choosing the Spanish terms that retrieved the Spanish pairs of the English documents retrieved for the English query. This approach is more effective than his previous ones, achieving up to 73.5% of monolingual performance.

From the discussion in Section 2.2, we know that a drawback of parallel corpora is that they are difficult to align properly. Even when the correct documents or sentences are aligned well, it is difficult to reconstruct the correct correspondences between a word in the source language and its translation in the target language. Parallel corpora tend also to have narrow coverage and may not yield the level of disambiguation necessary for a more general domain. The main disadvantage of parallel corpora is that they are a scarce resource.

Recall that comparable corpora are comprised of documents in different languages. Unlike parallel corpora, related documents are not translations of one another. Rather, they are documents written independently in different languages about the same topic or domain. Peters and Picchi [100] combine dictionary translation with comparable corpus analysis. They try to identify relationships between contexts in different languages under the assumption that translation equivalents will be found in related contexts. The method proceeds by identifying a source language term t to be translated and constructing a context window of n words around it. The significance of the correlation between each occurrence of t and its co-occurring words is measured by the Mutual Information Index [51]. The set A of the most significant co-occurring words is constructed and characterizes t. The words in A are translated via a bilingual machine readable dictionary, and the translations become the elements of set B. Windows in the target text for which terms from B occur significantly are considered to be contexts related to t. The most probable translation for t is chosen from these contexts. Although the technique looks promising, there have been no reports of its effectiveness for retrieval.

Work at ETH [113] has focused on using the SDA comparable corpora discussed in Section 2.2 to build similarity thesauri. SDA documents are aligned via manually-assigned descriptors. There are over 200 possible descriptors indicating characteristics such as topic, event type, date and country of origin. Aligned documents are then concatenated to generate a collection of multilingual documents. A thesaurus is generated in which co-occurrence forms the basis of the relationships between concepts. Replacement of query terms in one language with foreign language terms co-occurring with them in multilingual documents generates a translation effect. In other words, words co-occurring in the multilingual documents are rough translations of each other.

This method has been shown to be especially effective when the corpora are domain specific [114].

However, it is not clear that comparable corpora are easier to construct or align than are parallel document collections. As with parallel corpora, the question remains of what other disambiguation methods could be used in a more general context to augment these techniques.

Researchers at IBM [89] have had some success in applying the statistical MT approach of Brown to German query translation. Recall from Section 2.3.5.3 that this approach employs statistical analysis of a parallel corpus to generate language models to infer the most probable translation. In this study, researchers were able to achieve in the area of 70% monolingual effectiveness. A modified approach applied to a comparable corpus achieved 60% monolingual effectiveness. The limitations to this approach are primarily the dearth of parallel corpora and the complexity of comparable corpora alignment.

3.1.5 Bilingual Dictionary

Dictionary translation has been the starting point for other researchers [12, 13, 14, 101, 68]. Automatic machine readable dictionary (MRD) query translation leads to a drop in effectiveness of 40-60% below that of mono-lingual retrieval [68, 12]. This is due primarily to translation ambiguity. Statistical techniques which will be described later are an effective means of reducing ambiguity and its negative effects. Dictionaries like the other resources mentioned, may be proprietary, costly, or of limited availability, but are less so than aligned corpora.

MRD translations are typically performed via simple dictionary look-up. Translation is not used here in the sense of deep linguistic analysis. The terms of a query in one language are merely replaced with the dictionary definition of those terms in another language. This approach has been referred to as *lexical transfer*. The difficulty with dictionary translations is that they are ambiguous. The types ambiguities associated with translation are described in more detail in Chapter 5.

Hull's work has focused primarily on developing weighting models for ambiguity reduction. In [67], he describes a new weighted Boolean model that relies on conjunction to automatically disambiguate queries. One feature of the proposed model is that it enables the user to specify the importance of query terms by allowing user-specified belief parameters. Theoretical analyses of the model have been performed, but no reports of its effectiveness in practice have been reported.

Pirkola studies the effects of two factors on dictionary-based translations. The first is to tailor dictionary structure to aid disambiguation. The second is to add structure to queries to reduce ambiguity. He and Ballesteros [14] show independently that applying a synonym operator to translations is especially effective. The operator treats all target language translations for a single source language term as synonyms. Its effectiveness is due to a normalization for the variability in number of translation equivalents across source language words. This technique is discussed in more detail in Chapter 6, which describes our techniques for ambiguity reduction.

3.2 Summary and Discussion

Research in the area of cross-language information retrieval has focused mainly on methods for translating queries because full document translation for large collections is impractical. Methods for translation have included manual indexing and cognate matching, but have focused primarily on the use of machine translation techniques, the employment of parallel or comparable corpora, and dictionary translation.

Each of these approaches has limitations and disadvantages as described above. Manual indexing is not feasible for large collections. Cognate matching is only applicable to closely related languages. Machine translation systems are expensive to implement and modify, and require more information for accurate translation than a query typically contains. Aligned corpora are hard to come by and performance is dependent on how well the two corpora are aligned. Alignment and subsequent identification of term associations is not a trivial task. Finally, simple dictionary translations are ambiguous.

Ballesteros's work is discussed more thoroughly in Chapters 6 and 7 of this dissertation. It shows that a dictionary-based approach, augmented with statistical techniques for ambiguity reduction that are based on co-occurrence, is effective. First, the next chapter describes the experimental methodology for the work and then discusses methods used for evaluation.

CHAPTER 4

EVALUATION AND EXPERIMENTAL METHODOLOGY

This chapter discusses the experimental methodology and evaluation of the techniques employed in this dissertation. All experiments in this study were performed using the INQUERY information retrieval system which is discussed in Section 1.2.2.1. We begin by describing the test collections in Section 4.1. Section 4.2 describes the measures of effectiveness, followed by a discussion of our experimental methodology in Section 4.3.

4.1 Test Collections

Retrieval effectiveness is evaluated with test collections consisting of sets of queries and document collections for which we have relevance judgments. Relevance judgments tell us a priori which documents from a given collection are relevant to each query. We can then evaluate a system on its ability to retrieve those documents known to be relevant to a given query set. Evaluation experiments in this dissertation employed standard test collections including TREC-3, TREC-4 Spanish (ISM), TREC-6 Cross-language French (SDA French) and English (AP88-90). These test collections were generated for the annual Text Retrieval Conference administered by the National Institute of Standards and Technology (NIST). A brief description of each collection follows and Table 4.1 gives statistics about them.

- TREC-3: This is the test collection for the English ad hoc queries for TREC-3 [61]. It consists of 2.2 Gigabytes of English documents from the Associated Press

Newswire 1988-1989, Department of Energy abstracts, Federal Register 1988-1989, Wall Street Journal 1987-1992, and Ziff Davis Computer-Select Articles. The query set has fifty queries (151-200), of which we used twenty (151-162,164-171). There are approximately 167 relevant documents per query.

- TREC-4 Spanish (ISM): This is the test collection for the TREC-4 [62] Spanish ad hoc queries. It contains 208 Megabytes of Spanish documents from the Mexican newspaper, El Norte. The collection has twenty-five queries (26-50) of which we used twenty (26-45). There are approximately 91 relevant documents per query.

- TREC-6 Cross-language: This is the test collection for the Cross-language ad hoc queries for TREC-6 [98]. The full Cross-language collections contains English, French, and German text, however we employed only the English and French portion in this work. This portion of the collection consists of 748 Megabytes of English newswire documents from the Associated Press (AP) 1988-1990 and 250 Megabytes of French newswire documents from the Swiss news agency (SDA) 1988-1990. The newswire collections were chosen to overlap in time-frame to increase the likelihood of finding a greater number of relevant documents in all languages. Because the news tends to report current events, different news agencies will report on the same events and topics. This means that if an article written in English about a particular event is relevant, chances are high that an article in French about the same event will also be relevant. There were twenty-five queries formulated in English, Spanish, and French. For the transitive translation experiments, we used nineteen of those (1,2,4-7, 9-14, 16-21, and 23) for which relevant documents existed in both the English and French portions of the corpus. Queries 3,8, 15, 24, and 25 have no relevant English documents and queries 22, 24, 25 have no relevant French documents,

Table 4.1. Statistics on TREC collections used to evaluate cross-language retrieval. Stop words are not included in collection or query statistics and stop phrases are not included in query statistics.

Collection	Language	Size (GB)	Document Count	Terms per Doc	Query Count	Terms per Query	Relevant Docs per Query
TREC-3	English	2.2	741,856	260	20	10.75	167
TREC-4 ISM	Spanish	0.2	57,868	284	20	4.95	91
TREC-6 AP	English	0.75	242,918	458	21	7.8	59
TREC-6 SDA	French	0.25	141,656	185	20	9.1	57

so these queries were not evaluated. There are approximately 59 relevant English documents and 57 relevant French documents per query. This dataset is a new collection with pooled relevance judgments from thirteen retrieval systems. However, the preliminary nature of the data shouldn't greatly effect the outcome of our experiments.

The Cross-Language collection of TREC-6 was constructed in such a way as to contain equivalent queries formulated in English, French, and German. Equivalent queries in Spanish and Dutch were also provided. As mentioned above, the newswire documents were chosen from the same time frame. This was done to maximize the probability that although documents were written independently in different languages, they would describe the same events and thus increase the likelihood of finding relevant documents in all languages. With the exception of this collection, the others were constructed for monolingual or multilingual retrieval, that is, for retrieval with queries and documents in English or for retrieval with both queries and documents written in a language other than English.

Due to there being no available cross-language collections prior to TREC-6, for our earlier work we simulated one by manually translating the queries for each monolingual collection into another language. The English TREC-3 queries were manually translated to Spanish yielding a set of Spanish queries with relevance judgments for

English documents. Spanish TREC-4 queries were treated similarly. The manual translation of the Spanish queries was performed by a graduate student fluent in Spanish and whose native language is English. The manual translation of the English queries was performed by a graduate student fluent in English and whose native language is Spanish. Table 4.2 shows an example of a query and its manual translation. Queries constructed by manually translating an existing query set are referred to as *base* queries.

Table 4.2. Example TREC query before and after manual translation.

Original TREC Query	Relaciones económicas y comerciales de México con los paises asiáticos , por ejemplo Japon, China y Corea
Manual Translation	Economic and commercial relations between Mexico and the Asiatic countries, for example, Japan, China, and Korea

It is well-known that IR techniques can be sensitive to differences in retrieval variables including document length [116], query structure [64], and the language of queries and documents [47]. In an effort to generate results that are generalizable, we work with test collections having varying characteristics. Collection size ranges from 201 MB for TREC-4 Spanish to 2.2 GB for TREC-3 English. Average document length ranges from 185 to 458 words and average number of relevant documents ranges from 57 to 167. In addition, we apply the cross-language approach to retrieval between three languages: English, Spanish, and French.

4.2 Measures of Effectiveness

4.2.1 Recall and Precision

Recall and precision are two metrics typically calculated to evaluate an IR system. A system that is capable of retrieving all relevant documents would be said to have high recall and one in which most of the retrieved documents are relevant would be said to have high precision. More specifically, recall measures the ability of a system to retrieve all relevant items and is the proportion of relevant documents in the database actually retrieved by the system. It is given by

$$\frac{\#relevant\ items\ retrieved}{\#relevant\ items\ in\ collection}$$

Precision measures the ability of a system to retrieve only relevant items and is given by

$$\frac{\#relevant\ items\ retrieved}{total\ \#\ items\ retrieved}$$

These measures evaluate the quality of an unordered set of retrieved documents.

Most modern IR systems do not return a set of documents, rather they return a list of documents ranked by their probability of relevance to the query. To evaluate an ordered list of documents, precision can be plotted against recall after each retrieved document. Average performance over a set of queries, each possibly having a different number of relevant documents, is facilitated by interpolating over a set of standard recall levels from 0 to 1 at increments of 0.1. Precision at recall point 0.0 is the precision for the first relevant document in the ranked output. For the remaining recall levels, the rule applied is that the precision for a query at recall level l is the maximum precision obtained for the query at any actual recall level greater than or equal to l. The average of precision values across standard recall levels can be calculated as a single metric to measure the quality of the ranked output. This is known as 11-point average precision. To evaluate effectiveness of an IR system using a set of queries, the precision at a given recall level is taken as the average of the precision values for all queries at that recall level. Table 4.3 shows a recall-precision table for two techniques, A and B. For each technique, the table shows the total number of relevant documents for the query set, the total number of relevant retrieved, interpolated recall-precision averages and non-interpolated average precision. It also gives precision values at document cut-offs (e.g. after n documents have been retrieved where n = 5, 10, 15, documents and so on).

Table 4.3. Recall-Precision Table showing 11-point average precision, non-interpolated average precision, and precision at document cut-offs.

Technique	A	B
Relevant:	3255	3255
Rel Retrieved:	1479	1686
Interpolated Recall - Precision Averages:		
0.00	0.5676	0.7064
0.10	0.3636	0.4779
0.20	0.2907	0.3575
0.30	0.2226	0.2863
0.40	0.1940	0.2184
0.50	0.1713	0.1644
0.60	0.1426	0.1328
0.70	0.1183	0.1034
0.80	0.0831	0.0365
0.90	0.0355	0.0175
1.00	0.0000	0.0000
Avg precision (non-interpolated) over all rel docs		
	0.1792	0.2076
% Change:		15.9
Precision:		
5 docs:	0.3200	0.4900
10 docs:	0.3800	0.4800
15 docs:	0.3433	0.4500
20 docs:	0.3175	0.4225
30 docs:	0.3150	0.3867
100 docs:	0.2265	0.2675
200 docs:	0.1657	0.1965
500 docs:	0.1083	0.1253
1000 docs:	0.0739	0.0843

To avoid problems associated with interpolation, non-interpolated average precision over all relevant documents can be calculated. Rather than an average of the precision at standard recall levels, it is the average of the precision value obtained after each relevant document is retrieved. When no relevant document is retrieved, precision is assumed to be 0. This measure reflects the performance over all relevant documents and rewards a system for retrieving relevant documents at high ranks. Non-interpolated average precision is the measure reported for retrieval experiments described in this document.

Average precision is used as the basis of evaluation for all experiments. It is unrealistic to expect a user who may only be fluent in the query language to read and judge the relevance of many retrieved foreign language documents, so we also report precision at low recall levels. All work in this study was performed using the INQUERY information retrieval system. INQUERY is based on the Bayesian inference net model and is described briefly in Section 1.2.2.1 and in more detail in [123, 122, 20].

Monolingual system output provides a sense of the kind of effectiveness one could expect from a cross-language system if we were able to eliminate the confounding effects of "translation" ambiguity. For this reason, we also report the effectiveness of the cross-language approach as a percentage of monolingual effectiveness. More specifically, we take the ratio of the average precision for cross-language retrieval to that of monolingual retrieval for the same set of queries. Recall from Section 4.1 that this is possible because we have multiple manual translations of each query set.

4.2.2 Significance Tests

An IR experiment, like any other experiment, is affected by random error. This makes it difficult to say with any certainty that observing a difference in the effectiveness of two techniques is due to anything other than chance. For this reason, we perform hypothesis testing. If samples are small, a t-test [87] is only valid for normal distributions. Given that our query sets consist of only twenty queries and since the underlying population distributions for performance values such as average precision measures is not known, only weak statistical tests can be applied. The non-parametric significance tests that we apply are the Sign Test [65, 34] for paired data and the Wilcoxon Signed-Rank [65, 34] test.

The null hypothesis we test is the assumption that the techniques being compared are equally effective. The significance test does not allow us to conclude that one technique is better than another on the basis of performance differences. What it does allow us to test is whether the observed difference is statistically significant. In other words, the test calculates how often one would get a difference as large or larger than the one obtained if the difference were due to chance. If the difference would only occur by chance very rarely, we can reject the null hypothesis that there is no difference in performance between the two techniques.

The *significance level* or α is a measure of how unlikely it is that this difference would occur. The lower the significance level, the less likely observations are due to chance. In this thesis, we use $\alpha = 0.05$ or $\alpha = 0.01$. More specifically, for $\alpha = 0.05$, we can conclude that if chance is responsible for the observed difference, the difference would be obtained only 5% of the time or less. A similar statement can be made for $\alpha = 0.01$.

The sign test tests the hypothesis that the median of the difference between pairs is zero. The test statistic is the number of positive differences. If the null hypothesis is true, there should be approximately equal numbers of positive and negative differences.

One disadvantage of the sign test is that is does not take into account the magnitude of the difference between pairs. The Wilcoxon Signed-Rank does take this into account. The test is similar to the sign test in that it tests for the median difference between pairs to be zero. However, it does so by sorting the absolute values of the differences, assigning ranks to them, and then finding the sum of the ranks of the positive differences. If the null hypothesis is true, the sum of the ranks of the positive differences should be about the same as that for the negative differences.

4.3 Experimental Methodology

Ultimately, we are interested in evaluating the ability of our cross-language approach to address the ambiguity associated with translation. If cross-language ambiguity can be sufficiently reduced, we could expect cross-language effectiveness to be comparable to that of monolingual retrieval. Recall from Section 1.4 that we take a machine readable dictionary (MRD) approach to the cross-language retrieval problem. This section describes the use of the MRD in the query translation process. Techniques for ambiguity reduction and their application will be discussed in Chapters 5 and 6.

4.3.1 Query Processing and Indexing

The queries provided by TREC are written in such a way that they more resemble essay questions than user queries. They observe proper grammar, punctuation, and tend to be wordy. They typically consist of several fields. Table 4.4 gives an example of two TREC queries.

Table 4.4. Two queries in the form in which they are distributed by TREC. The first is the English version of TREC-6 query 1. The second is TREC-3 query 151.

```
<num> Number: CL11
<E-title> Organic Cotton
<E-desc> Description:
Documents containing information on production and uses of organic cotton.
<E-narr> Narrative:
A relevant document will contain information on production of and uses
for organic cotton that also provide ecological benefits to the
soil.
```

```
<num> Number: 151
<title> Topic: Coping with overcrowded prisons
<desc> Description:
The document will provide information on jail and prison overcrowding
and how inmates are forced to cope with those conditions; or it will
reveal plans to relieve the overcrowded condition.
<narr> Narrative:
A relevant document will describe scenes of overcrowding that have
become all too common in jails and prisons around the country. The
document will identify how inmates are forced to cope with those
overcrowded conditions, and/or what the Correctional System is doing,
or planning to do, to alleviate the crowded condition.
```

Because queries are long and wordy, they are automatically pre-processed to make them more resemble an actual user query. The first step in this process is to remove all fields other than *Title* and *Description*. What remains after step one is a shorter query, however it still contains common phrases which do not provide any content specific information. The second step is to remove these *stop-phrases.* Example stop-phrases are “A relevant document will” and “The document will discuss/describe”. When part-of-speech tagging is employed for phrase recognition, this takes place prior to stop-phrase removal.

The indices for both Spanish and English documents were generated after removal of stop-words and consist of individual word stems. The stemmers employed were those included in the INQUERY retrieval system. The English stemmer called KSTEM [77] is based on rules for inflectional and derivational morphology. The Spanish stemmer is based on the Porter stemming algorithm [102] and inflectional morphology rules for Spanish. The environment for French, however, differs slightly from that for English or Spanish. We do not have a French stemmer so we employ the XEROX Finite-State Morphological Processor [132] to simulate the effect of stemming by expanding query terms with inflectional variants.

4.3.2 Word-by-word MRD Translation

MRD translations are performed after simple morphological processing of query terms to remove most plural word forms and to replace verb forms with their infinitive form. We wrote a Spanish morphological processor in *flex* [97] that is based on the rules of pluralization and verb conjugation for Spanish. Recall that we do not use the word “translation” in the sense of deep linguistic analysis. The words of a query in one language are merely replaced with the dictionary definition of those words in another language. Morphological processing for French was performed via the on-line XEROX morphological analyzer [132].

Given that dictionary definitions tend to list several senses each having one or more related meanings, translations are ambiguous. To reduce ambiguity, we initially chose one of two replacement methods. The first method is to replace a query term with only those words or *translation equivalents* listed for the first sense of the term. We assume that the first sense listed is also the most frequent. The negative effect of this is that some relevant meanings will be lost. This approach is referred to as the *sense1* method. The second method is to tag each query term with its part-of-speech (POS) and to replace a term with only that part of the dictionary entry corresponding to its POS.

Part-of-speech tagging is an automatic means of identifying a word's grammatical category. The input to a tagger is a natural language sentence and the tagger assigns each word in the sentence its part-of-speech. Taggers are readily available now and have been developed for many languages. They have primarily been based on probabilistic or rule-based techniques [15, 19, 31] with both approaches yielding high levels of accuracy.

Our empirical results (Table 6.15) suggest that neither replacement approach is more effective than simple replacement with all translation equivalents when other ambiguity reduction techniques are employed. Words which were not found in the dictionary were added to the new query without translation. The Collins [28] English/Spanish bilingual MRD was used for the Spanish-English and English-Spanish translations. Translations between French and English were performed using the Collins French/English bilingual MRD. Spanish-French translations were performed via the Larousse Spanish/French bilingual dictionary [82].

4.3.2.1 Query Expansion

4.4 Summary

The work in this dissertation explores the viability of a dictionary approach to cross-language retrieval. The basic experimental methodology is to employ machine readable dictionaries to generate approximate query translations and then to apply ambiguity reduction strategies to those translations. Retrieval is performed on corpora for which we have relevance judgments. In other words, we know which documents in the collection are relevant to each query. Effectiveness of this dictionary-based approach is evaluated by analyzing the ability of the system to retrieve all relevant documents (recall) and to retrieve them at the top of the ranking (precision). In addition, we perform monolingual retrieval on the same queries as it indicates how well the system should perform without the confounding effects of ambiguity. We then compare average precision of cross-language retrieval to that of monolingual retrieval to get a sense of how effective our disambiguation strategies are. Chapters 6 and 7 describe in detail our ambiguity reduction strategies and their effectiveness. However, the next chapter first describes the types of ambiguities that arise in a cross-language environment. It then discusses approaches to ambiguity reduction and provides a foundation for the discussion of our approaches.

CHAPTER 5

AMBIGUITY AND TRANSLATION ERROR

Our goal for the cross-language task is to generate an approximate translation of a user query. The preceding chapter describes the resources that can be employed to solve to this task. These include lexical resources such as word lists and dictionaries of varying degrees of complexity, and parallel or comparable corpora, as well as language analysis tools from natural language processing (NLP) and machine translation (MT). Regardless of the types of resources employed, each approach to the CLIR task must find ways to address the ambiguity problem, of which there are two types: within-language (WIL) ambiguity and between-language (BL) ambiguity .

WIL ambiguity has to do with identifying the meaning of a particular linguistic item, such as a word, phrase, or sentence. Disambiguation techniques typically focus on individual word senses. Addressing WIL amounts to inferring the intended meaning of the query.

BL ambiguity is related to issues regarding the translation of a linguistic item from one language into another. This is not a trivial task and is complicated by many factors. First, the mapping of concepts between languages is not one to one. There are for example, concepts in one language that do not exist in another. Second, even if the conceptual framework between two languages is fairly close, a non-ambiguous concept in one language may be lexicalized in many ways in another language. The difficulty becomes identifying which of the translation equivalents is appropriate for the given context.

WIL and BL ambiguities are discussed in Sections 5.1 and 5.2. Section 5.3 discusses approaches for reducing ambiguity and provides the background for Chapter 6, which describes our own approaches to ambiguity reduction in a cross-language retrieval environment.

5.1 Within-language Ambiguity

When a word, phrase, or sentence is open to more than one interpretation, it is said to be ambiguous. Within-language ambiguity relates to determining the meaning conveyed by a text. This type of ambiguity can be semantic or syntactic in nature.

Syntactic ambiguity occurs when the structure of a sentence leads to more than one possible interpretation. For example, "Heating fuels can be dangerous" may be interpreted to mean either that *applying heat to fuels is dangerous* or that *fuels used for heating are dangerous.* These differ in syntactic structure. Ambiguity in this case would be resolved if a different choice of the verb "be" were chosen. "Heating fuels is dangerous" implies the first interpretation given above, while "Heating fuels are dangerous" implies the latter.

Semantic ambiguity occurs when the same word has more than one meaning. This can occur by several different means. First, a word can have one or more grammatical categories or senses. For example, *row* can be a noun (objects in a straight line) or a verb (propel by oars) and *square* can be either a noun (rectangle with four equal sides), a verb (to make square), or adjective (precisely constructed or aligned).

Even if a word does not have multiple senses, it can still have multiple meanings. The word *boxer* can be a fighter, a person who packs things in boxes, or a breed of dog. A *program* can be either the print-out of the order and features of a performance or a sequence of coded instructions for a computer. These words are known as homonyms and people typically use contextual or real-world information to identify which meaning of the word is intended. Given the sentence "The boxer's paw was bitten by the fox." and our knowledge that paws are not a characteristic of humans, it is not difficult for a person to understand that "boxer" refers to an instance of that breed of dog. However as discussed in Section 2.3.2, representing and employing this kind of knowledge by computer is non-trivial for a couple of reasons. First, it is difficult to identify a priori which pieces of contextual information will eventually be useful. Second, due to the tremendous amount of potentially useful information, it is not currently feasible to effectively store and exploit it.

Homographs are words for which the same issues arise. They have have the same spelling, but different meanings and possibly different pronunciations. The difference in pronunciation has no effect on text analysis, however text processing can not take advantage of auditory clues to aid disambiguation as a human might.

5.2 Between-language Ambiguity

Between-language ambiguity arises when trying to translate text expressed in one language into another language while preserving the essence of the original text. BL ambiguities are related to differences in the way that ideas are expressed across languages. These differences may be lexical or structural. These types of ambiguities are discussed in more detail below.

5.2.1 Lexical Ambiguity

Lexical ambiguities occur when there are several choices of target equivalent. Consider the translation of the word "branch" into Spanish. If one means a division of a larger organization, it should be translated to *sucursal*, to *ramo* or *ramo* in the context of a tree branch, or to *bifurcación* if a branch of a computer program.

Languages may also have lexical differences based on the granularity of concept classification or the underlying state of affairs. One may, for example, make a much finer distinction than another as in the following case. I have two brothers, one born before me and one born after. To distinguish between them, I may refer to one as my *older* brother and to the other as my *younger* brother, but English uses the same lexical item "brother" to describe both. This is not the case in Chinese, which uses different words; the word "didi" refers to a younger brother and "gege" refers to an older brother.

Distinctions between concept classifications can also be based on the underlying state of affairs. The translation to Spanish of the English word *know* is based on the kind of knowledge one has. Given "I know her.", *know* should be translated using the verb *conocer*, as is appropriate for knowledge of people. In the sentence "I know her address.", *know* should be translated using *saber* as it is associated with knowledge of facts.

Especially challenging is the case where there are no words in one language to represent a concept in another language. The Japanese word *mokusatsu* [40] is an example of a word for which there is no English equivalent. This word describes the Japanese practice of using time breaks in business negotiations to psych-out the other party. In the middle of a discussion, Japanese negotiators will simply stop talking. They may get up and leave the room or may sit silently with their eyes closed for short periods of time. It literally means "killing with silence" and according to De Mente refers to the idea that the other party's case will die in the vacuum of silence. Words such as these are typically not found in any dictionary, leading to the out-of-vocabulary (OOV) problem. The difficulty in dealing with OOV terms is that there is no resource that provides a direct mapping of the terms to equivalent translations in the target language.

Each of the problems described above poses a problem for translation. In the most extreme case, there may be no direct translation or no lexical database or dictionary containing an appropriate translation equivalent. Similar problems arise from structural ambiguities such as those encountered in the translation of phrases or *collocations.* Collocations are phrases that can not be translated word-by-word. Structural ambiguities are described in the following section.

5.2.2 Structural Ambiguities

Structural ambiguities arise from differences in the way different languages express the same concept. A concept may be represented by one word in one language and be represented by several words or a phrase in yet another. For example, the English word "rowboat" is translated as the Spanish phrase "bote de remos". If there is a dictionary entry for the word, then translating from the word to a phrase is not a problem. It is typically more difficult to automatically translate a phrase. First, the system must be able to correctly identify phrases. After the phrase has been identified, some means of distinguishing the correct translation must be found if the phrase is not listed in an available dictionary. Table 5.1 shows some Spanish phrases with their English translations as given by the Collins Spanish/English MRD. When no dictionary translation is available, *compositional phrases* can be translated word-by-word if the correct translation equivalent for each of its constituent words can be accurately identified. Other types of phrases are not translatable word-by-word.

Spanish Phrase	English Equivalent
apio nabo	celeriac
Barba Azul	Bluebeard
recin casado	newly-wed
recin nacido	newborn

Table 5.1. Spanish phrases that translate to a single English word.

Metaphorical and idiomatic expressions are typically more difficult to translate than other types of phrases. Because their structure may not be predictable or consistent, they are more difficult to identify. Their translations may also map to one word in another language. The correct translation of many multi-word metaphors and idioms is another multi-word expression, however they are not easier to translate because they are not translatable word-by-word. Furthermore, inclusion in dictionaries of these multi-word constructs is far less prevalent than other types of phrases. Table 5.2 gives examples of idiomatic and metaphorical expressions with their literal translations and true meanings or corresponding English expressions.

Expression	Literal Meaning	True Meaning
Hace mucho frío.	It makes much cold.	It is very cold.
Tiene hambre.	S/he has hunger.	S/he is hungry.
El hijo de la gata, ratones mata.	The son of the cat kills mice.	Like father, like son.
se bourrer la gueule	to stuff the mouth	to get drunk

Table 5.2. Some idiomatic phrases, their littoral translations, and their true meanings.

5.3 Resolving Ambiguity

In the course of our daily communications, we are often confronted with the ambiguities of natural language. However, humans have very little difficulty resolving semantic ambiguity. In the case of machine disambiguation, syntax may be applied when sentences are grammatically correct. In some cases, cues such as number agreement between subject and object pairs can also be applied. Consider the example from Section 5.1, "Heating fuels can be dangerous". We pointed out that this ambiguity could be resolved syntactically by re-stating the sentence with some form of the verb *to be*. Given the following two sentences, it is clear that in "Heating fuels *is* dangerous." "dangerous" is a predicate of the verbal noun "heating" while in "Heating fuels *are* dangerous." "dangerous" is a predicate of "fuels" and heating modifies fuels. However, syntactic clues are not helpful for other types of ambiguities. Take for example, the word *bank*, which may be a financial institution, earth at the edge of river, or a container for coins. Humans bring to bear real-world knowledge and MT systems typically try to perform the same level of disambiguation. Section 2.3.2

presents details about the level of linguistic analyses that are most often done in an MT environment in an effort to attain that goal. It also discusses why this is no trivial task for a computer, thus shedding light on the reasons why MT systems typically fall short of their goal.

Much research has been done to develop automatic word disambiguation methods. Unlike many of the MT approaches, this section focuses on techniques that do not rely upon linguistic analysis. Instead, the approaches presented below rely in part on the concept of term co-occurrence as described in Section 1.2.3.3. Recall that the basic idea behind term co-occurrence is that we can learn something about the meaning of a word by looking at the words that occur in the same context. The goal of this section is not to review all approaches to disambiguation, but rather to describe some of the main approaches. This will provide a background for discussions of our own approaches to disambiguation given in chapter 6.

5.3.1 Dictionary Approaches

A number of techniques [86, 57, 110, 111, 77, 127, 104, 137] for resolving ambiguities rely upon engineered lexical resources such as dictionaries, thesauri, or term lists. The basic approach is to analyze lexical information and the structure imposed on it, to try to identify finer-grained relationships between individual words.

For many applications, lexical disambiguation has been assumed to be helpful, but has typically not been shown to be so. Several researchers have employed information found in machine-readable resources to the task. Lesk [86] employed word overlap for lexical disambiguation. His goal was to try to guess the correct sense of a word for a given context. The approach proceeds by comparing the definitions of each sense of the ambiguous word with the definitions of words that co-occur with it. The sense whose definition has the most overlaps with co-occurring word definitions is chosen as the correct sense. Limited tests on small amounts of text yielded 40-70% accuracy.

Errors were related to the way in which dictionaries are constructed and to data size. Definitions frequently give many examples of word usage which would often throw the algorithm off. In addition, there was often not enough overlap either because some definitions are too short or because even related definitions may use different vocabulary to describe the same topic.

Guthrie et al.[57] suggested using co-occurrence statistics to disambiguate words found in the Longman's Dictionary of Contemporary English (LDOCE) [83]. The assumption was that words co-occurring with a particular sense of a word generate a "core context" or "neighborhood" for that sense. The neighborhood of a word was constructed via co-occurrence analysis of the definitions in which the word appeared. The LDOCE marks each sense with a subject field code that indicates the subject area. Definitions were grouped by LDOCE subject headings in order to generate subject-dependent neighborhoods. A neighborhood could be expanded incrementally by the addition of new words co-occurring with words from the neighborhood derived on the previous iteration. A different neighborhood was constructed for each context because a word tends to occur with different words in different contexts. Lexical disambiguation was performed by measuring the overlap between a word's neighborhood and the text to be disambiguated. No reports of the direct application of this approach were reported.

Other researchers have examined the effects of lexical ambiguity directly on information retrieval. Krovetz and Croft [78] show through query analysis that ambiguity is typically not a big problem in IR because of a natural disambiguation effect of a query. Those documents containing many query words tend to be more highly ranked than those containing fewer query words. When an ambiguous word occurs with many other query words in the text of a document, it is because the document uses the ambiguous word in the same sense as the query. They suggest that ambiguity will be a bigger problem for more diverse databases and for high-precision searches that are based upon the importance of a single concept, especially when there is little overlap between document and query.

Later, Krovetz [77] attempted to resolve ambiguity indirectly via stemming. Recall from Section 1.2.1 that stemming is a technique for conflating words that refer to the same concept. Krovetz's stemmer is based on the belief that conflation should be based on meaning and not necessarily on words sharing the same root.

In this work, a morphological stemmer was built to take advantage of several types of information. For example, irregular morphology can be used to identify meaning. Krovetz gives an example of the word *antennae* which can be the plural of the type of antenna that is associated with insects but not a television. In addition, suffixes only attach to roots with a particular part-of-speech so this information can be used to differentiate between homographs. A look-up table was used to enforce some of these types of rules. The use of the stemmer yielded significant performance improvements, particularly when documents were short. In addition, Krovetz ran overlap experiments similar to those of Lesk. However, he used overlap to identify relationships between morphological variants by comparing definitions of their different senses. This approach yielded a rate of success between 80-96%.

Like Krovetz, Voorhees [127] used conflation as an approach to disambiguation in an IR context. Information contained in the WordNet [90] lexical database was the basis for generating conflation classes. WordNet is organized at the highest level into four parts, one each for nouns, verbs, adjectives, and adverbs. Each part is then organized into *synsets* which group words into sets of strict synonyms. This means that if a word occurs in more than one synset, each synset contains a different sense of the word.

Voorhees' study focused on disambiguating nouns. It was based on the notion that a given context determines the correct senses for a set of individually ambiguous words co-occurring in that context. To exploit this idea automatically the approach assigns to each word sense, contextual categories derived from synset analysis. For each category, the number of words in a text that have senses belonging to that category is counted. The categories having the largest counts are assumed to contain the correct senses of the ambiguous words. To test the effectiveness of sense resolution, a comparison of retrieval is performed between documents indexed using the synsets of disambiguated words versus those indexed using word stems. Effectiveness for the disambiguation technique was much lower than for word stems alone. This was due primarily to the difficulty in disambiguating queries. Since queries are shorter, they have significantly less context to use for disambiguation. This creates a mismatch between the word senses selected for disambiguation of documents and those selected for queries. The author notes that retrieval results suggest the approach would not be effective for word sense disambiguation. It is also suggested that the relationships represented in WordNet do not contain enough information to make distinctions at a granularity that is fine enough for the task. However, no evaluation of the approach for merely selecting the correct sense of an ambiguous word was performed.

Sanderson's [110] work on ambiguity in IR supports that of Krovetz and Croft, showing that ambiguity is not a big problem unless queries are short or there is little overlap between words in the query and the document. He also shows that if a disambiguator is employed, the method must be very accurate in order to have a positive effect on performance. Sanderson's approach was to introduce various levels of ambiguity into document collections by the creation of pseudowords. These pseudowords were generated by concatenating some number of adjacent words in the text to be indexed. After introducing ambiguity, disambiguation at various levels of accuracy was performed before retrieval and then retrieval effectiveness was evaluated. The study showed that disambiguation at only 75% accuracy was far worse than without disambiguation for queries of varying lengths. Retrieval after disambiguation with 90% accuracy was nearly as effective as retrieval without disambiguation, but showed improvement with queries of length 1-2. This supports the results given above for Voorhees' work in which discrepancies between the level of disambiguation between queries and documents led to poor retrieval effectiveness.

The previous approaches rely primarily on analysis of precoded knowledge contained in resources such as MRD's, thesauri, or semantic networks to implicitly identify lexical relationships. Although the approaches yield some good results, they are often confounded by the structure imposed by the builder of the lexical resource. Other work relies on analysis of the information in aligned corpora to automatically infer lexical relationships. Such techniques are the topic of the next section.

5.3.2 Aligned Corpus Disambiguation

Aligned corpora consist of a collection of sets of related documents. Parallel corpora contain a set of documents and their translations in one or more other languages. The structure in these collections typically consists of merely identifying which documents are translations of one another. In a comparable corpus (described more fully in Section 2.2), documents are not direct translations of one another but are related by topic or context. Analysis of either type of paired documents can be used to infer the most likely translations of terms or concepts between languages in the corpus. Many approaches to disambiguation rely upon the use of parallel corpora to identify the most probable translations [22, 117, 79, 125, 32, 48, 74].

Brown, et al. [22] employed a parallel corpus in English and French to disambiguate individual words. The approach relies on Brown's earlier work described in Section 2.3.5.3 and is based on developing a translation model which estimates the probability that an English sentence, E, is the translation of a French sentence, F. The model is a function of two items. The first is a probabilistic model of the English language and the the second is the process of translating from French to English. The model is derived from analysis of a sentence alignment of the parallel corpus.

For each sentence pair, the translation model incorporates the idea that each English word acts independently to generate one or more French words. A mapping between an English word and a French word it produces is said to be an *alignment* (A). Given a particular English sentence, E, the probability that a French sentence, F, is E's translation is calculated by summing over all the probabilities of all possible alignments and is given by

$$P(E|F) = \sum_{A} Pr(F, A|E).$$

The most probable alignment (*Viterbi alignment*) for a sentence pair is inferred via a Viterbi estimation. The probability that a particular French word is the translation of an English word is then estimated using the Viterbi alignments for an aligned corpus.

In this study, over 1 million short sentences from the Canadian Hansard corpus were employed as the training corpus, producing more than 12 million word pair alignments. Given the context in which a word w appeared, w was assigned two senses based on the co-occurrence of some other word within w's context. This co-occurring word was referred to as an *informant.* The idea is that given some arbitrary text window in which w appears, locating an informant of w in the same window will tell us something about the context in which w should be translated. To resolve ambiguity in the translation of a French word *f*, the set of English words aligned with f (E_f) is partitioned into two groups. The set of possible informants of f (I_f) is partitioned similarly. The partitioning algorithm splits E_f and I_f such that there is a maximal mutual information between English and French partitions. The resulting English partitions define the translation equivalents of the two senses assigned to *f.* Each French partition maps to one of these senses and its members are used as informants to identify the sense of f to be translated. The technique worked reasonably well in distinguishing between the two partitions or senses generated by the algorithm. However it was not employed in a more realistic setting in which ambiguous word have more than two possible senses.

The technique described by Smadja, et al focuses on the disambiguation of collocations and is the most general of the aligned corpora techniques. It includes translation of single words and phrases of different types and has flexible rules about proximity. For example, a phrase can be a sequence of adjacent terms or may contain an arbitrary number of other words. Their system, Champollion, translates collocations given a parallel corpus aligned at the sentence level. Collocations are phrases that can not be translated word-by-word. For a given source language phrase, Champollion uses the Dice coefficient to identify target language terms that are highly correlated with it. The words are then combined in a systematic, iterative manner to produce a translation of the source language phrase. All pairs of words are considered and those that are highly correlated to the source phrase are identified and passed to the next stage. Triplets are then produced by adding a highly correlated word to a highly correlated pair and these highly correlated triplets are passed to the next stage. The process is repeated until no more highly correlated combinations of words are found. Word order is determined by looking at examples in the corpus.

Evaluation was done using a "collocation compiler", XTRACT, to automatically produce several lists of source phrases. Three sets of three-hundred medium-frequency collocations were extracted from parallel text from the Canadian parliamentary proceedings. Human evaluation showed that the system correctly translated 68-75% of the source phrases. The best results are given when the training and test corpus are the same.

Although the system is relatively accurate, it has several disadvantages. First, it is too slow to use on the fly. Running with 220 MB of aligned sentences (640,000 pairs), it takes 1-2 min. to translate a phrase. Second, infrequent phrases can not be translated with the method.

5.3.3 Disambiguation via Unaligned Corpora

The approaches mentioned above identify source language words or phrases and rely upon the use of aligned corpora to identify the most probable target language translations. Acknowledging the disadvantages of manually constructed resources and aligned corpora, other researchers [33, 76, 112] have focused on exploiting the information implicit in unaligned corpora.

Kraaij and Hiemstra [76] used co-occurrence frequency for term translation with some success during the TREC-6 [129] evaluations. They disambiguate by looking at co-occurrence statistics of adjacent query words. However, no details of the technique were reported.

Dagan, et al. [33] employs a co-occurrence method for target word selection. Their technique first generates phrases,*word1 word2*, from words paired via syntactic relationships e.g. subject-verb. Only those phrases left after applying deterministic syntactic constraints are disambiguated. The method proceeds as follows. Given two sets A and B, containing the translations of *word1* and *word2* respectively, all possible translations for the pair are generated by taking the cross product of the sets. Selection is made via a statistical model based on the ratio of the frequency of co-occurrence for one alternative versus the frequency of co-occurrence of all other alternatives. Overlap due to terms occurring in more than one phrase is exploited to aid the disambiguation of other phrases. The method shows some promise, with 75-92% precision on a small number of phrases. However, there have been no reports of its use in a retrieval environment and the conditions under which the method was tested are unrealistic for a general retrieval domain. First, translation equivalents were manually filtered to remove rare translations. Second, the technique relies on syntactic analysis and therefore requires syntactically correct queries. Finally, it was found to be applicable to only a small percentage of the phrases tested.

5.4 Summary and Discussion

This chapter describes the different ways that the intended meaning of language objects can be obscured by lexical imprecision or syntactic structure. Each of these factors not only affects our ability to understand the import of a piece of text but can also confound automatic methods. As humans, we are familiar with the ambiguous nature of language. We are constantly applying our own knowledge of the world to aid us in ascertaining the meaning of the messages we receive. The same degree of information can neither be encoded nor exploited efficiently by machine, so automatic methods can employ approaches that fall somewhere between two extremes.

On one end of the spectrum, they simulate to some degree the processes by which humans disambiguate language. That is, they use processing which combines knowledge about grammar, morphology, syntax, and semantics manually constructed for the task, to try to determine a parsed representation of the text that is unambiguous. Analysis at this level of granularity is extremely difficult to perform correctly and efficiently by machine. This is the approach taken by many machine translation systems and is the subject of Section 2.3. Due to the manual effort to construct resources and the computational cost of employing this approach, most MT systems are limited to a sublanguage or small domain of discourse. Despite these limitations, the output of an MT system must typically be hand-edited before it is distributed. When the goal is to generate exact, syntactically correct representations of a text in other languages, many are satisfied with an automatic system that falls short. This is because the system may lessen the burden on humans who would otherwise perform the entire task alone. However, other tasks do not require such a lofty goal. In cases such as these, the level of performance does not justify the cost of linguistic analysis. For this reason, many researchers look to general approaches that give reasonable performance at a lower cost.

At this end of the spectrum the focus is primarily on lexical and/or phrasal disambiguation. Section 5.3 of this chapter discusses many such approaches. The resources they employ range from highly structured items such as dictionaries and thesauri, to aligned corpora, where the structure amounts to explicitly identifying relationsips between documents or parts of documents, to unaligned corpora where there is no explicit relationship between items in the collection. In the case of dictionaries and thesauri, techniques analyze the structure imposed on lexical items to identify relationships. When aligned corpora are employed, fine-grained relationships between lexical items are inferred via analysis of relationships at the document level. For unaligned corpora, relationships must be inferred via statistical analysis of word co-occurrence.

Although they may not perform deep linguistic analysis, the approaches discussed in this section are not without cost. The availability of each type of resource varies as does the degree of processing necessary to employ them effectively. Dictionaries and unaligned corpora are the most readily available resources. Aligned corpora, typically generated under special circumstances, are not abundant and are generally proprietary. With the exception of unaligned corpora, each of the resources requires some degree of preprocessing. When employing dictionaries and word lists, the processing may be no more than filtering mark-up intended for human analysis. However, in aligned corpora relationships are typically identified at the document level. For this reason, they tend to require more complex processing such as the generation of more fine-grained alignments between words or sentences. This is not a trivial task and the effectiveness of disambiguation will be dependent upon the quality of the alignments. A more detailed discussion of the costs of multilingual resources is given in Chapter 2.

In the chapter that follows, we describe our own approach to disambiguation in a cross-language retrieval environment. The approach employs the most readily available resources: machine readable dictionaries and unaligned corpora. We begin by analyzing the effects of ambiguity on a dictionary-based approach. We then describe our approaches to ambiguity resolution and present their effectiveness for cross-language retrieval. Finally, we show that the level of linguistic analysis performed by MT systems is not necessary for effective CLIR.

CHAPTER 6

RESOLVING TRANSLATION AMBIGUITY

Chapter 5 describes the variety of ways that ambiguity can be manifested thus making it difficult to discern the intended meaning of a text object. The goal of this research is to find methods for reducing ambiguity that do not rely on scarce resources such as parallel corpora. Bilingual machine-readable dictionaries (MRDs), which are more prevalent than parallel texts, are a good alternative. In this chapter, we develop a dictionary-based approach to cross-language retrieval. We begin by examining the the discrepancy in performance between a monolingual and a cross-language retrieval system. We analyze the sources of translation ambiguity and their impact on retrieval performance. We then describe statistical techniques and show that they can significantly reduce the impact of ambiguity and bring the effectiveness of cross-language retrieval near that of monolingual retrieval.

6.1 Translation Ambiguity and Retrieval Effectiveness

In the experiments described in this chapter, we analyze the effectiveness of our cross-language approach by comparing it to a monolingual system. The main idea is to simulate a cross-language environment such that the efficacy of a query translation approach can be evaluated and compared to monolingual retrieval. The goal is to generate an approximate translation in which the gist of the original query is preserved. In order to do this, we need a set of queries in a source language, a collection of documents in a target language, and relevance judgments. Recall that relevance judgments indicate which documents are relevant to each query. In addition to a set of queries in one or more source languages, we also have the original target language queries. We then evaluate the effectiveness of a translation approach by measuring the ability of the translated queries to retrieve relevant documents. Performance comparisons are made on the same query and document collections, with the only difference being that the cross-language system must first translate its queries to the language of the monolingual system. The natural approach to automatic translation via machine readable dictionary is simple word-by-word (WBW) replacement. More specifically, each word in the source language is replaced by its translation equivalents in the target language. For a detailed description of our experimental methodology, refer back to Chapter 4.

6.1.1 Simple WBW Dictionary Translation and Ambiguity

It stands to reason given all that we know about translation ambiguity and automatic methods, that cross-language retrieval via simple dictionary replacement would be less effective than monolingual retrieval on the same queries. In fact, Ballesteros and Croft [12] and Hull and Greffenstett [68] show that word-by-word query translations yield cross-language retrieval effectiveness that is less than half as effective as monolingual retrieval. There are several problems with a simple dictionary approach that lead to this drop in effectiveness.

First, word-by-word translations are inherently ambiguous. For each head-word, a dictionary will list several parts-of-speech each having one or more related meanings. The resulting translation will contain many incorrect translation equivalents that are used in different contexts. For example, the Spanish translation of “retard” could be either “retardar” or “retrasar” both meaning “to slow”, however the former is used in the context of growth and the latter is used in the context of progress. A query will generally refer to one context. Adding both translations would increase the ambiguity of the new target language query since it may retrieve documents related

Term	Meaning	MRD Translation
mundo	world	world, people, society, secular life
conocer	know	know, to know about, understand, meet, get to know, to become acquainted with
country	país	país, patria, campo, región, tierra

Table 6.1. Examples of terms, their meanings in particular queries, and their MRD word-by-word translation.

to both contexts. Table 6.1 gives some examples of the dictionary entries for the first sense of several words.

Second, queries often contain multi-word concepts that lose their intended meaning when translated word-by-word. Consider the Spanish phrase *oso de peluche*, meaning *teddy bear*. The dictionary lists (*bear, braggart, bully*) and (*felt, plush*) as the translation equivalents for oso and peluche respectively (*de* is a preposition meaning *of*). It is not possible to reconstruct the correct English translation of the phrase via a word-by-word approach. Table 6.2 gives examples of other phrases, their meanings, and their word-by-word translations via the Collins English-Spanish dictionary. There are *compositional* phrases for which the correct translation can be derived word-by-word, however it is still non-trivial to select the appropriate equivalent for each word. Although the difference between a phrasal translation and a WBW translation may be subtle, it can have a large negative impact on effectiveness [12]. More details about the impact of phrasal translations are given in the following section.

Phrase	Meaning	WBW translation
cifras del costo	cost figures	amount of the cost
fondos de inversión	mutual funds	fund of investment
marina de guerra	navy	navy of war

Table 6.2. Examples of phrases, their meanings and their word-by-word translations.

Finally, there are some words that can not be translated via a dictionary. Dictionaries typically contain vocabulary used in a variety of settings. This broad coverage makes them applicable for translating queries covering a wide variety of topics. However, coverage is generally not deep enough to include many domain specific words or specialized terminology. Query words that can not be translated via the dictionary are referred to as *out-of-vocabulary* (OOV) words.

Figure 6.1 shows recall-precision curves for monolingual retrieval and cross-language retrieval via word-by-word replacement on the TREC-3 query set. Recall from Section 4.1 that TREC-3 and TREC-4 consisted of English data and Spanish data, respectively. In order to simulate a cross-language environment, their query sets were manually translated into Spanish/English. In the experiments described below, these manual translations are referred to as *base* queries. To perform the retrieval for the graph, cross-language queries were translated in the following way. Each cross-language query word was translated by performing a dictionary look-up and then replacing the query word with all definitions listed for the first sense of the dictionary entry. This graph clearly illustrates the degradation in effectiveness for this simple cross-language approach.

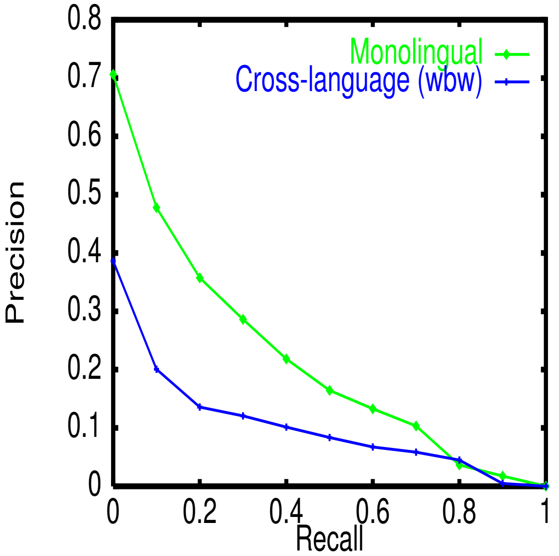

Figure 6.1. Recall-precision graph comparing monolingual retrieval effectiveness to that of cross-language retrieval via simple word-by-word (wbw) replacement.

To find the extent to which each of the causes of ambiguity is responsible for drops in cross-language effectiveness, two additional translations of the base TREC-4 queries were generated by hand. The first was a word-by-word translation in which we chose the one best target language word to replace each source query word. The second was created similarly but with phrasal translations where appropriate. Results show that performance does improve with the refinement of each query set, as illustrated in figure 6.2. It compares the effectiveness as measured by average precision of each manual translation and the automatic translation to that of the monolingual (target language) query set. The first bar gives the percentage of monolingual effectiveness achieved by each translation. The percentage drop in effectiveness of the automatically translated query set below that of the monolingual set is considered to be translation error. The bar on the right shows the percentage of this "error" that each factor is responsible for. The transfer of senses inappropriate to the query accounts for up to 26% of the loss of effectiveness while phrase loss accounts for up to 31%. An additional 23% of the loss can be attributed to the exclusion of acronyms and specialized terminology. A similar experiment was also performed on the TREC-3 base queries. In this case, we found that word and phrase ambiguity is responsible for 51% and 24% of the error, respectively. These results are supported by the work of Hull and Grefenstette [68] who performed similar query translations from French to English.

Table 6.3 gives term statistics for the TREC-3 and TREC-4 base queries and their translations. The first column is the query set name. SE-BASE and ES-BASE queries are the manual translations of the original TREC-4 and TREC-3 queries, respectively. SE-1st and ES-1st refer to the automatic wbw translation of the base queries, using only the first sense of a dictionary definition. The second column is the mean number of terms per query, the third is the mean number of terms that were not found in the MRD, the fourth is the mean number of terms returned from the MRD per translated

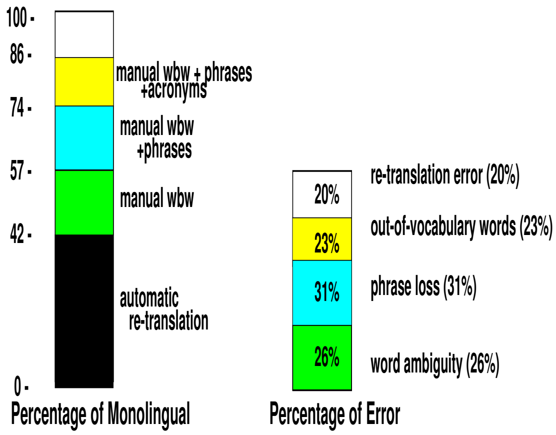

Figure 6.2. Effect of translation ambiguity on effectiveness as measured by average precision.

Query-set	Query-terms	Undefined	Terms per translation	Orig. terms recovered
Original Spanish	5.35(3.33)	N/A	N/A	N/A
SE-BASE	4.95(3.35)	13	N/A	N/A
SE-1st	12(33.4)	N/A	4.05	2.1(1.19)[3.15(1.93)])
Original English	10.6(16.54)	N/A	N/A	N/A
ES-BASE	10.6(21.14)	3	N/A	N/A
ES-1st	33.3(149.2)	N/A	3.09	4.4(5.14)[6.05(9.65)]

Table 6.3. Query term statistics for TREC-3 and TREC-4 queries after stop words and stop-phrase removal.

query term, and the last is the mean number of original query terms recovered after translation. Variances are shown in parentheses. Given that all queries are stemmed (see Section 1.2.1 for a discussion of stemming) prior to retrieval, we also show (in square brackets) the statistics for recovered terms after stemming.

There is little difference between the length of the original queries and that of their manual translations. Automatic translations of the ES-BASE and SE-BASE queries yield a recovery of 58% and 59% of the original query terms, respectively. However the translations differ in that 82% of the ES-1st terms are not original query terms, so may be erroneous. This number is higher than for SE-1st, with 64% of its terms being possibly erroneous. This may explain the higher percentage of translation error attributable to word ambiguity for the TREC-3 queries (51%) than for the TREC-4 queries (26%).

In addition, some error is attributable to those words for which no correct translation was found in the dictionary (OOV words). The remaining loss in effectiveness can probably be attributed to less well specified queries and to ambiguity introduced through the original manual translation. Manual re-translations of the base queries were performed to test this hypothesis. We found that the manual re-translations yielded 90% and 98% of the effectiveness of the original TREC-4 and TREC-3 queries, respectively. This suggests that some of the drop in effectiveness may in fact be attributable to error associated with the translation that generated the base queries. Tables 6.4 and 6.5 show four representations of several TREC-4 queries. The first is the original query, second is the base query (manual translation of the original), third is the automatic WBW translation of the base, and finally we show the manual re-translation of the base query. Tables and 6.6 and and 6.7 show similar representations of several TREC-3 queries.

TREC-4 q26	relaciones económicas y comerciales de México con los paises europeos
Base	the economic and comercial relations between Mexico and the european countries
Manual of Base	relaciones económicas y comerciales entre México y los paises europeos
Auto of Base	(económico equitativo rentable) comercial (narración relato relación)(Méjico México)europeo (país patria campo región tierra)

Table 6.4. Four query representations for TREC-4 query 26: Original, Base (manual translation), Manual re-translation of Base, and MRD WBW translation of base.

Although there is some variation in construction across queries, for example the number of out-of-vocabulary words or phrases, results are consistent for all of our query sets. They show a difference in effectiveness of between 40-60% when comparing cross-language retrieval via simple dictionary translation to monolingual retrieval.

TREC-4 q31	medidas tomadas por el gobierno mexicano para resolver la disputa con los rebeldes zapatistas en el estado de Chiapas
Base	the methods taken by the mexican government for resolving the dispute with the zapatista rebells in the state of Chiapas
Manual of Base	medidas que he estado tomado por el gobierno mexicano para resolver la controversia con los rebeldes zapatistas en el estado de chiapas
Auto of Base	(método sistema procedimiento) of (mejicano mexicano)(gobierno administración Estado régimen) resolución (disputa discusión altercado conflicto laboral contencioso) zapatista rebelde estado condición Chiapas
TREC-4 q32	la importancia de las Naciones Unidas (NU) para México
Base	the importance of the United Nations for Mexico
Manual of Base	la importancia de la Organización de las Naciones Unidas para México ONU
Auto of Base	importancia Nations (Méjico México)

Table 6.5. Four query representations for two TREC-4 queries: Original, Base (manual translation), Manual re-translation of Base, and MRD WBW translation of base.

TREC-3 q151	information about jail and prison overcrowding and how inmates are forced to cope with those conditions, reveal plans to relieve the overcrowded conditions
Base	información sobre el hacinamiento en las cárceles y las prisiones y como los internos se ven forzados a afrontar estas condiciones, revelará los programas para paliar el hacinamiento
Manual of Base	information about overcrowding in jails and prisons and how the inmates are forced to deal with these conditions will reveal the programs for alleviating the overcrowding
Auto of Base	(heaping stacking crowding overcrowding accumulation)(prison jail)prison(internal interior inside) (forced compulsory) (to bring face to face)(nature condition temperament character)(to reveal to disclose to betray show to give away)(programme scheme plan) (to palliate mitigate alleviate to relieve to lessen cushion to diminish)(heaping stacking crowding overcrowding accumulation)

Table 6.6. Four query representations for TREC-3 query 152: Original, Base (manual translation), Manual re-translation of Base, and MRD WBW translation of base.

Phrases account for between 31-36% of this drop in retrieval effectiveness and single

TREC-3 q153	insurance coverage for long term care confinements
Base	cobertura de seguro por confinamiento a largo plazo
Manual of Base	health insurance for long term confinement
Auto of Base	(cover covering)(safe secure)(bordering adjoining adjacence contiguousness)(long long lengthy too long)(time period term deadline time limit date expiry date date)
TREC-3 q155	implications of the decision of Right Wing Christian Fundamentalist Groups to use the political process to further their goals.
Base	implicaciones de la decisión de los grupos fundamentalistas cristianos de derecha de usar el proceso político para promover su objetivo.
Manual of Base	implications of the decision by Christian fundamentalists to use the political process to promote their objectives
Auto of Base	contradiction (decision judgement) (group cluster clump group) fundamentalist Christian (right hand right side right-hand side) (to use make use of to wear) process political (to promote advance further to promote to pioneer to sponsor to begin set on foot get moving to bring) objective

Table 6.7. Four query representations for two TREC-3 queries: Original, Base (manual translation), Manual re-translation of Base, and MRD WBW translation of base.

word ambiguity due to irrelevant translation equivalents and OOV words account for up to 50%.

6.1.2 Dictionary Translation of Phrases

In the previous section we explain why dictionary methods for cross-language information retrieval give performance below that of monolingual retrieval. Failure to translate multi-term concepts as phrases greatly reduces the effectiveness of dictionary translation. Our hypothesis is that automatically identifying phrases and defining them as such will improve effectiveness. In experiments where query phrases were manually translated [12], effectiveness improved by up to 25% over automatic word-by-word (WBW) query translation. To test this hypothesis, we compare performance of automatically translated queries both with and without phrasal identification and translation.

Individual words were translated automatically as described in Section 4.3.2.Briefly, queries are first POS tagged. Spanish and English queries were tagged with the BBN part-of-speech tagger[15]. The TreeTagger system [71] was used for tagging French. (Recall that tagging takes place prior to stop phrase removal.) Sequences of nouns and adjective-noun pairs were taken to be phrases. Query words in the source language are replaced with the dictionary definition of those terms in the target language.

Since the true meaning of a collocation or phrase is often impossible to generate via a word-by-word translation, we avoided this problem when possible by translating multi-word concepts using information gathered from the MRDs. Phrasal translations were performed using a database built from information on phrases and word usage contained in the Collins or Larousse MRD. During phrase translation, the database is searched for source language phrases that are contained in the query. A hit returns the target language translation of the source language phrase. If more than one translation is found, each of them is added to the query. This allowed the replacement of a source phrase with its multi-term representation in the target language. When a phrasal translation could not be found in the dictionary, it was translated word-by-word. Words that are not found in the dictionary are added to the new query without translation. Table 6.8 gives some examples of phrasal translations extracted from the Collins English/Spanish Dictionary.

Table 6.9 gives recall-precision values for wbw translation of the TREC-4 queries without and with dictionary translation of phrases. Results suggest that in this case, phrasal translation does not improve effectiveness. It gives average precision values for a baseline of automatic WBW translation vs automatic WBW with phrasal translation. A closer look at individual queries reveals that phrasal translation is not ineffective, but that results are sensitive to poor translations. This supports the

Table 6.8. Phrasal translations.

Phrase	Translation
United Nations	Naciones Unidas Organización de las Naciones Unidas
trade agreement	convenio comercial
South Africa	Unión Sudafricana Africa del Sur
member country	los países miembros los países afiliados los países participantes los países pertenecientes

disambiguation results of Voorhees, Krovetz, and Sanderson discussed in Chapter 5. Their results suggest that unless disambiguation methods are highly accurate, attempts to disambiguate may further degrade effectiveness.

	WBW	Phrasal
Avg Precision	0.0823	0.0826

Table 6.9. Average precision of WBW vs phrasal translation. (TREC-4 ISM Spanish corpus, queries 26-45, and English-> Spanish translation)

For example, average precision drops 40% below a baseline of automatic WBW translation for query SP30 when phrasal translations are included. However, the problem for this query is that "sports program" is translated as "emisiòn deportiva" meaning "televised sports program". When the poor phrasal translation is replaced with a WBW translation, effectiveness improves three-fold (+150% over the baseline). Table 6.10 shows 5 representations of SP30: Original, base (manual translation), automatic WBW translation, automatic phrasal + WBW translation, and automatic WBW translation after replacing the "poor" phrasal translation via WBW translation. Parentheses enclose recognized phrases and brackets enclose phrasal translations. Results for the last three query representations are given in Table 6.11.

Original TREC-4 Spanish Query	programas y intercambios deportivos entre México y los Estados Unidos
Manual Translation to English	(Sports programs) and (exchange programs) between Mexico and the (United States)
MRD Translation to Spanish	deporte caza deporte juego diversión víctima juguete programs canje cambio intercambio programs Méjico México estado condición
MRD Word-by-word and Phrasal Translations to Spanish	[emisión deportiva] cambio canje intercambio programs [Estados Unidos][el coloso del norte] [Estados Unidos de América] Méjico México
MRD Word-by-Word and Corrected Phrasal Translations to Spanish	deporte caza deporte juego diversión víctima juguete programs cambio canje intercambio programs [Estados Unidos] [el coloso del norte] [Estados Unidos de América] Méjico México

Table 6.10. Five query representations for SP30: Original, base (manual translation), MRD translation, MRD WBW + phrasal translation of base, MRD WBW + phrasal translations of base - poor phrasal translations.

	WBW	Phrasal	Corrected Phrasal
Avg	0.0244	0.0148	0.0610
% Change:		-39.3	150.3

Table 6.11. Average precision for WBW vs two different phrasal translations for TREC-4 query SP30. (TREC-4 ISM Spanish corpus and English-> Spanish translation)

These results suggest that well-translated phrases can greatly improve effectiveness, but that poor translations may degrade effectiveness. Translation accuracy may be more important for phrases than for single words.

6.2 Ambiguity Resolution

Despite the negative effect of translation ambiguity on cross-language retrieval effectiveness, ambiguity reduction techniques can be applied to significantly improve effectiveness. Most of the techniques we apply are based on statistical analysis of word co-occurrence, but simple techniques based on syntax and query structure are also employed.

The following sections describe techniques for reducing the effects of ambiguity associated with simple dictionary translation. As described in the preceding chapter, when ambiguity reduction is not employed, cross-language retrieval via simple MRD translation is only 40-60% as effective as monolingual retrieval. However when the techniques described below are applied, the effects of ambiguity can be greatly reduced.

6.2.1 Synonyms and Part-of-Speech

If a source language term has more than one target language equivalent, its translation will be ambiguous. Here we demonstrate the disambiguating effect of two simple techniques. First, we reduce the effect of irrelevant term translations by building structured queries with the synonym operator (#syn). Second, we reduce the number of target language equivalents by replacing each source term with only those equivalents corresponding to a term's part-of-speech. Queries are replaced by all translation equivalents unless POS is applied. In this set of experiments, all phrases were translated word-by-word.

There are two factors related to erroneous word translations that reduce effectiveness. First, because dictionaries often include archaic usages for head-words, query translations can be dominated by these rarely used equivalents. Second, query words having the greatest number of translation equivalents tend to be given greater importance. This is an artifact of the way in which document relevance is assessed. The words that a query and a document have in common serve as the basis for inferring the likelihood that they are related. Thus a query word replaced with ten translations will have five times as many chances to match document terms than a query word replaced with only two translations.

We measure the importance or discriminating power of a word in a collection of documents by a *belief score* based on two types of term frequency, *tf-score* and *idf-score*. Tf-score reflects the within document frequency of a term while idf-score is inversely proportional to the number of documents in the collection in which the term occurs. For example, a document containing the word *apple* with reasonable frequency is a good indication that the document is about apples. However, if the word *apple* appears in many documents across the collection, it will not have much ability to discriminate between relevant and non-relevant documents about apples. The score for *apple* in this particular document would get credit for a high tf-score, but would be penalized for occurring frequently throughout the corpus (low idf-score). Infrequent or rare terms have higher idf scores, so tend to have higher belief values.

The synonym operator treats occurrences of all words within it as occurrences of a single pseudo-term whose term frequency is the sum of the frequencies for each word in the operator. This de-emphasizes infrequent words and has a disambiguating effect. If the synonym operator is not used, infrequent translations get more weight than more frequent translations due to their higher idf. The correct translation of a query term is generally *not* an infrequently used word, so in most cases this approach is effective.

In addition, the synonym operator reduces ambiguity by normalizing for the variance in number of translation equivalents across query terms. Without the synonym operator, query words with many translation equivalents would get a higher weight than words with fewer translations. This is because an alternative to the synonym operator is to give each translation equal weight. In a two word query t_{s1} t_{s2} in which query terms t_{s1} and t_{s2} have one and five translation equivalents respectively, the resulting target language query, t_{t11} t_{t21} t_{t22} t_{t23} t_{t24} t_{t25}, essentially treats the concept described by t_{s2} as five times as important as t_{s1}. Application of the synonym operator has the effect of treating occurrences of all translations of a word as an occurrence of a single concept, therefore normalizing for this variance.

When queries are expressed in well-formed sentences, POS tagging can be employed. This enables the replacement of each query word with only those translation equivalents corresponding to its correct POS. Part-of-speech tagging may not eliminate the transfer of archaic translation equivalents, but it can reduce ambiguity by reducing the number of erroneous terms. In our experiments, English and Spanish queries were tagged with the BBN part-of-speech tagger[15]. The TreeTagger system [71] was used for tagging French. The synonym operator is wrapped around all translation equivalents for a translated query word. For example, the Spanish translation of the first sense of *country* is *país, patria, campo, región, tierra.* The translated query would thus contain #syn(país, patria, campo, región, tierra).

The synonym operator is much more effective in general than is the application of POS. This is good news since queries are rarely expressed in syntactically correct sentences. The synonym operator yields improvements in effectiveness of greater than 45% over simple word-by-word translation alone, while POS disambiguation yields improvements of up to 22% for short queries. POS is less effective for long queries since the number of erroneous terms may not be sufficiently reduced to yield significant improvements in effectiveness.

Table 6.12 shows the positive effect on average precision for both techniques on short queries. Table 6.13 shows their effects on long queries. In each case, column one corresponds to a word-by-word translation (WBW) of all queries with no attempt at disambiguation. Column two shows the effect of the synonym operator on WBW. Column three shows a word-by-word translation using only POS to disambiguate. The last column combines the disambiguation effects of POS tagging and the use of the synonym operator.

The synonym operator is much more effective than is part-of-speech, with the former primarily affecting precision and the latter primarily affecting recall. POS is moderately effective. Combining the two techniques is most effective and improves

Query	WBW	SYN	POS	POS+SYN
Relevant Docs	1247	1247	1247	1247
Retrieved	379	390	667	737
Avg.Prec.	0.1234	0.1784	0.1504	0.2331
% Change:		44.6	21.9	89.0
Precision at:				
5 docs:	0.2286	0.2762	0.3048	0.3619
10 docs:	0.2286	0.2381	0.3000	0.3286
20 docs:	0.1929	0.2190	0.2476	0.3095
30 docs:	0.1667	0.1968	0.2286	0.2810
100 docs:	0.0786	0.1129	0.1362	0.1705

Table 6.12. Average precision and number of relevant documents retrieved for word-by-word translation, word-by-word translation augmented by POS disambiguation, synonym operator disambiguation, and word-by-word translation augmented by POS and synonym operator disambiguation of short queries. (TREC-6 AP English corpus, Spanish queries 1,2,4-7, 9-14, 16-24, and Spanish-> English translations.)

both precision and recall. However, empirical results given in Table 6.15 show that simple replacement with all translation equivalents is as effective as employing POS when both phrasal translation and synonym disambiguation is employed. It is good news that POS is not necessary since POS-tagging only works well for well-formed sentences.

For the queries where Spanish was translated to French, the effect of the synonym operator is significant at the level $p << 0.01$. Results are shown in Table 6.14. The effect is not statistically significant for short queries translating Spanish to English nor is application of POS alone on any query set. However, combining both the synonym operator and POS does give statistically significant results for the long Spanish to English translations.

The synonym operator is more effective when applied to longer queries. This may be true for the following reason. The number of translation equivalents will be higher for a longer query by virtue of the fact that there are more terms to translate. The translation of the long queries contains more inappropriate equivalents overall than the translation of the short queries. The synonym operator is also more effective when there is a greater number of translation equivalents per query term. Moreover, given

Query	WBW	SYN	POS	POS+SYN
Relevant Docs	3255	3255	3255	3255
Retrieved	409	1008	368	996
Avg.Prec.	0.0385	0.1530	0.0369	0.1529
% change		297.2	-4.2	297.0
Precision at:				
5 docs:	0.1400	0.4300	0.1000	0.4300
10 docs:	0.1150	0.4000	0.1000	0.3950
20 docs:	0.1150	0.3400	0.1075	0.3400
30 docs:	0.0983	0.3083	0.0883	0.3050
100 docs:	0.0690	0.2070	0.0680	0.2040

Table 6.13. Average precision and number of relevant documents retrieved for word-by-word translation, word-by-word translation augmented by POS disambiguation, synonym operator disambiguation, and word-by-word translation augmented by POS and synonym operator disambiguation of long queries. (TREC-3 English corpus, queries 151-162,164-171, and Spanish->English translations.)

that most words have only two parts-of-speech, any effect of disambiguating via POS is overwhelmed by the number of translation equivalents per query term.

Table 6.16 gives query term statistics including query length after translation with and without POS disambiguation for each query set. Note that the number of original query terms recovered after translation is essentially the same whether or not POS disambiguation is applied. Queries are also 10-15% shorter with POS disambiguation, however, the percentage of possibly erroneous terms is still 90% or greater. For the CSF queries which are shorter and for which the dictionary lists fewer translation equivalents per term, the percentage of possibly erroneous terms is only 73%.

6.2.2 Word Co-occurrence, Aligned and Unaligned Corpora

Words take their meaning from the context in which they are used and this fact can be exploited to boost the effectiveness of a query. In absence of any other information, if someone speaks of “the bank” we are unable to determine whether the reference is to the financial sense of the word or to the land bordering a body of water. However if one indicates that a deposit of money must be made, it can be inferred that the reference is to the financial sense. Analysis of word co-occurrence allows us to make

Query	WBW	SYN	POS	POS+SYN
Relevant Docs	1098	1098	1098	1098
Retrieved	522	586	533	595
Avg.Prec.	0.1818	0.2028	0.1864	0.2028
% change		11.5	2.5	11.5
Precision at:				
5 docs:	0.3400	0.3600	0.3600	0.3600
10 docs:	0.2800	0.2950	0.3000	0.2950
20 docs:	0.2175	0.2425	0.2175	0.2300
30 docs:	0.1833	0.2133	0.1867	0.2067
100 docs:	0.1195	0.1300	0.1225	0.1340

Table 6.14. Average precision and number of relevant documents retrieved for word-by-word translation, word-by-word translation augmented by POS disambiguation, synonym operator disambiguation, and word-by-word translation augmented by POS and synonym operator disambiguation of short queries translated from Spanish to French. (TREC-6 SDA French corpus, Spanish queries 1,2,4-7, 9-14, 16-24, and Spanish-> French translations.)

assumptions about the intended meanings of query words and thus reduce the effects of ambiguity.

One means of exploiting word co-occurrence is via query expansion (described in Section 1.2.3.1) where the results of previous retrievals are employed to improve the effectiveness of subsequent retrievals. The approach is to modify the query by adding words co-occurring with query terms in documents known or assumed to be relevant. This expansion with related words focuses the query and improves effectiveness.

In the cross-language environment where queries may contain many erroneous terms, application of query expansion is based on two assumptions. The first is that related terms will tend to co-occur in documents while unrelated terms will tend not to. The second, is that the documents containing the related terms will occur at the top of the ranking. If these assumptions are correct, then we hypothesize that we can treat ambiguity as an instance of vocabulary mismatch and improve effectiveness via expansion. Our work ([12, 13, 14]) shows that expansion of cross-language queries is effective at two stages of the translation process: prior to dictionary translation and after dictionary translation. Sections 6.2.2.1 and 6.2.2.2 present query expansion

Query	WBW	SYN	POS	POS+SYN	SYN+PHR	POS+SYN+PHR
Relevant Docs	1247	1247	1247	1247	1247	1247
Retrieved	379	390	667	737	881	878
Avg.Prec.	0.1234	0.1784	0.1504	0.2331	0.2934	0.2944
% Change:		44.6	21.9	89.0	137.8	138.6
Precision at:						
5 docs:	0.2286	0.2762	0.3048	0.3619	0.4095	0.3905
10 docs:	0.2286	0.2381	0.3000	0.3286	0.4095	0.3714
20 docs:	0.1929	0.2190	0.2476	0.3095	0.3833	0.3738
30 docs:	0.1667	0.1968	0.2286	0.2810	0.3444	0.3413
100 docs:	0.0786	0.1129	0.1362	0.1705	0.2329	0.2329

Table 6.15. Average precision and number of relevant documents retrieved for word-by-word translation, word-by-word translation augmented by POS disambiguation, synonym operator disambiguation, word-by-word translation augmented by POS and synonym operator disambiguation, word-by-word plus phrase dictionary translation augmented by synonym dismabiguation, and word-by-word plus phrase dictionary translation augmented by synonym and POS dismabiguation of short queries. (TREC-6 AP English corpus, Spanish queries 1,2,4-7, 9-14, 16-24, and Spanish-> English translations.)

results for pre- and post-translation expansion, respectively. Section 6.2.2.3 presents retrieval results for combining these two approaches.

Expansion prior to translation (pre-translation expansion) creates a stronger base query for translation by adding terms that emphasize query concepts. Consider the English query "Programs for suppressing or limiting epidemics in Mexico" which translates to "Programas para reprimir o limitar epidemias en México". Pre-translation

Qry	Qry	Database		Qry	Defs	Undef.	Orig Terms
Set	Lang	Lang	POS	Length	per Term	Terms	Recovered
CSE	Spanish	English	no	54.67(738.98)	6.58(3.5)	4	4.62(6.81)
CSE	Spanish	English	yes	46.52(447.49)	5.62(21.43)	4	4.67(6.60)
CSF	Spanish	French	no	16.67(33.28)	2.11(3.39)	6	4.19(7.01)
CSF	Spanish	French	yes	15.05(61.7)	1.92(2.28)	6	4.14(7.17)
MSE	Spanish	English	no	88.2(1641.76)	7.68(58.62)	16	5.1(4.39)
MSE	Spanish	English	yes	78.1(1274.39)	6.73(42.18)	16	4.9(3.99)

Table 6.16. Mean (variance) statistics for TREC-6 English (CSE), TREC-6 French (CSF), and TREC-3 English (MSE) query sets: terms per query, definitions per term, undefined terms, number of original query terms recovered after translation.

expansion leads to the addition of *cholera, disease, health* and *epidemiologist* thus strengthening the intent of the original query. The subsequent query translation will contain more terms related to the information need, thus will mitigate the effects of erroneous term translations.

When post-translation expansion is applied to the query above, *morbo, morbosidad,* and *contagio* are examples of the words added to the query. In English, these translate to (morbidity, sickness rate), (morbidity, morbidness, unhealthiness, sick rate) and (infection, contagion, corruption, taint), respectively. Post-translation query expansion has the effect of further reducing the effects of erroneous term translations by adding more context specific terms.

We first apply local feedback [9] to each query set either prior to (pre-translation) or after (post-translation) MRD translation. Pre-translation expansion should improve results by adding terms that emphasize query concepts. Post-translation expansion is expected to decrease ambiguity by de-emphasizing inappropriate translation equivalents. In addition, both pre- and post-translation expansion is performed to see whether the positive effects of each alone are additive. Fig. 6.3 is a flow chart of query processing for these experiments.

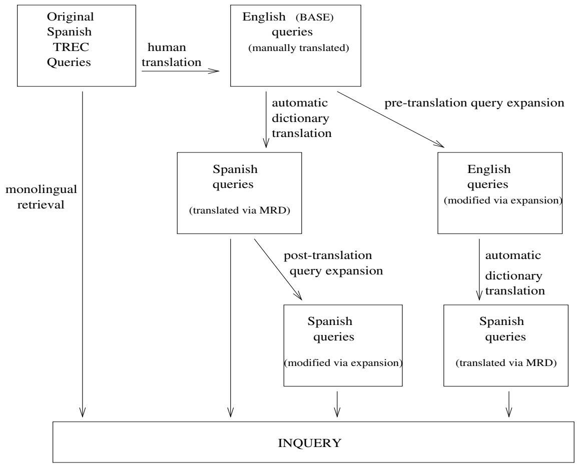
Original Spanish TREC Queries
human translation
English (BASE) queries (manually translated)
automatic dictionary translation
pre-translation query expansion
monolingual retrieval
Spanish queries (translated via MRD)
English queries (modified via expansion)
post-translation query expansion
automatic dictionary translation
Spanish queries (modified via expansion)
Spanish queries (translated via MRD)
INQUERY

Figure 6.3. Flow chart of query processing.

We then compare the effectiveness of Local Context Analysis (LCA) to that of local feedback. Recall that Local Context Analysis is a modification of local feedback and has been shown to be more effective. It differs from local feedback in two ways. First, rather than analyzing entire documents, it expands the query with words from the most highly ranked passages. Second, the more frequently a word co-occurs with query terms, the higher it will be ranked. However, it also penalizes words that occur frequently throughout the corpus. Translations are performed as described in Chapter 4. When possible, an attempt is made to translate phrases using the phrase dictionary.

Post-translation expansion terms are extracted from the databases being searched as would be done in a monolingual retrieval environment. However, this is not possible in the case of pre-translation expansion. We are trying to evaluate the ability of our system to retrieve documents in one language given a query in another language. For example, assume we want to retrieve English documents in response to a Spanish query. Pre-translation expansion requires the addition of *Spanish* terms to the source (Spanish) query. In this case, the database to be searched is in English, so will not be applicable in the pre-translation expansion procedure. We need training data, in this example a Spanish database, from which to extract pre-translation expansion terms. In general, it might not be clear which database to use for pre-translation expansion. However many databases are specialized, which is good for queries related to those specific areas. In the following experiments, training data for the Spanish and English pre-translation expansions are the 208 MB TREC-4 ISM Spanish collection and the 301 MB TREC-4 San Jose Mercury News (SJMN) English collection, respectively.

In this section, we also present a technique based on co-occurrence statistics from unaligned corpora that can be used to reduce the ambiguity associated with phrasal and term translation. Our hypothesis is that the correct translations of query terms will co-occur as part of a sub-language and that incorrect translations will tend not to co-occur. This information could be used to translate compositional phrases, thus reducing the ambiguity associated with word-by-word translation. This method is then combined with other techniques for reducing ambiguity and more than 90% monolingual effectiveness is achieved. Additionally, we propose that disambiguation methods using unaligned corpora can be as effective as those using parallel or comparable corpora. The details of the parallel corpus method and the co-occurrence method are given in sections 6.2.2.4 and 6.2.2.5. Finally, a comparison of the co-occurrence method with parallel corpus and machine translation techniques is done to show that good retrieval effectiveness can be achieved without complex resources.

The experiments in this section were evaluated on the 21 Spanish queries from the TREC-6 cross-language data. Spanish (*source* language) queries were translated to English (*target* language). Table 6.17 gives sample queries and their correct translations. Co-occurrence statistics were collected from either the 748 MB AP88-90 or the 2.1 GB TREC VOL12 corpus.

Queries were processed in the following way. First, queries were tagged by a part-of-speech (POS) tagger. Sequences of nouns and adjective-noun pairs were taken to be phrases. Automatic translations were performed by translating phrases as multi-term concepts when possible and individual terms word-by-word. MRD word-by-word translations were performed as described in Section 4.3.2. Term translations were disambiguated via part-of-speech and the synonym operator as described in Section 6.2.1.

Phrasal translations were also performed as described in Section 6.1.2 using information on phrases and word usage contained in the Collins MRD. When a phrase could not be defined using this information, the remaining phrase terms were translated in one of two ways. Terms were translated word-by-word followed by parallel corpus disambiguation (PLC) described in section 6.2.2.4, or they were translated as multi-term concepts using the co-occurrence method (CO) described in section

Original TREC Query	Caso Waldeheim. Razones de la controversia que rodea las acciones de Waldheim durante la Segunda Guerra Mundial.
Manual Translation to English	Waldheim Case. Reasons for the controversy surrounding the actions of Waldheim during the Second World War.
Original TREC Query	Educación sexual. El uso de la educación sexual para combatir el SIDA.
Manual Translation to English	Sex Education. The use of sex education to combat AIDS.
Original TREC Query	Comida rápida en Europa. Qué tan exitosa ha sido la expansión de concesiones americanas en Europa?
Manual Translation to English	Fast food in Europe. How successful is the spread of American fast food franchises in Europe?
Original TREC Query	An alleged illegality committed by any entity seeking a contract on behalf of the U.S. Military Forces
Manual Translation to Spanish	Una presunta ilegalidad cometida por alguna entidad que busca un contrato en representacin de las fuerzas armadas de los estados unidos EEUU
Original TREC Query	Middle-East Peace Process. What is the attitude of the Arab countries toward the Peace Process in the Middle East?
Manual Translation to French	Processus de paix au Moyen-Orient. Quelle est l’attitude des pays arabes l’gard du processus de paix au Moyen-Orient?

Table 6.17. Five TREC-6 queries with their manual translations.

6.2.2.5. PLC disambiguates terms using the entire query as context, while the CO method typically only uses the context of a phrasal unit. All CO experiments were run with a text window size of 250 terms, which is approximately the length of a document. Section 6.2.2.5, also compares the ability of the CO method with that of the phrase dictionary alone for translating phrases. The types of phrases translated and the effectiveness of the methods are given. Section 6.2.2.6 compares disambiguation of term translations via CO with disambiguation via PLC. We also compare the effectiveness of CO and PLC for reducing the error caused by failure to translate phrases as multi-term concepts.

Section 6.2.3 reports on the effectiveness of combining disambiguation methods described in this section with query expansion techniques described in Section 6.2.2.Query expansion before or after automatic translation via MRD significantly reduces translation error. Recall that pre-translation expansion creates a stronger base for translation and improves precision. Expansion after MRD translation introduces terms that de-emphasize irrelevant translations to reduce ambiguity and improve recall. Combining pre- and post-translation expansion increases both precision and recall. Improvement appears to be due to the removal of error caused by the addition of extraneous terms via the translation process. In the following experiments, query expansion is done via Local Context Analysis (LCA).Training data for the pre-translation LCA experiments consisted of the documents in the 208 MB El Norte (ISM) database from the TREC collection.

6.2.2.1 Pre-translation Expansion

We expected pre-translation expansion to be more effective with shorter queries. Fewer query terms could mean fewer content bearing terms, which may yield a translation that is swamped by irrelevant words. Tables 6.18 and 6.19 show how ambiguity can reduce query effectiveness and how pre-translation feedback can reduce that ambiguity. In each example there are five representations of the same query. The first is the original TREC-4 query, the second is the manual translation (base), the third is the MRD translation of the base query, the fourth is the base query after pre-translation feedback expansion, and the last is the MRD translation of the latter. Words in parentheses were returned as multiple translations for one term. Terms in brackets were added by feedback.

Original TREC Query	relaciones económicas y comerciales de México con los paises asiáticos, por ejemplo Japon, China y Corea
Manual translation to English	economic and commercial relations between Mexico and the Asiatic countries, for example, Japan, China, and Korea
MRD translation to Spanish	(económico equitativo rentable)comercial (narración relato relación)(Méjico México)(asiático)(país patria campo región tierra) el Japón China Corea
English pre-translation query expansion	economic commercial relations mexico japan korea china [korean nuclear south north]
MRD translation of pre-translation expansion query	(económico equitativo rentable)(comercial)(narración relato relación)(mexico)(laca japonesa)(porcelana loza)(korea)(korean)(nuclear)(sur mediodía)(norte)

Table 6.18. Five query representations for SP28: original, manual translation to English, automatic MRD translation to Spanish, English pre-translation feedback, and MRD translation of the pre-translation query to Spanish.

Original TREC Query	programas para reprimir o limitar epidemias en México
Manual translation to English	programs for suppressing or limiting epidemics in Mexico
MRD translation to Spanish	(programs)(controlador mayoritario)(restrictivo) (epidémico)(Méjico México)
English pre-translation query expansion	programs controlling limiting epidemics mexico [epidemic cholera disease health]
MRD translation to Spanish of pre-translation expansion query	(programs)(controlador mayoritario)(restrictivo) (epidémico)(mexico)(epidémico)(cólera)(enfermedad morbo dolencia mal)(salud sanidad higiene)

Table 6.19. Five query representations for SP43: original, manual translation to English, automatic MRD translation to Spanish, English pre-translation feedback, and MRD translation of pre-translation feedback query to Spanish.

Performance of query SP28 gets worse with each translation. The problem with the first translation is that although all the original query terms are included, the query seems to get swamped by inappropriate word definitions. This is also a problem with pre-translation feedback. The problem is exacerbated because feedback returns all terms in lowercase form, which is an artifact of tokenizing/indexing. Consequently, dictionary lookup fails to find proper nouns or instead finds their common noun definition: e.g., china is translated to "porcelana, loza" which mean "porcelain" and "china plate" respectively. This latter error can be minimized by ensuring that proper nouns retain capitalization. Note that this is less of a problem when translating from Spanish to English since fewer proper nouns are capitalized: e.g., the translation of Australian is australiano.

The word by word translation of query SP43 also suffers from the problem described above. However, the pre-translation feedback improves performance considerably. The inclusion of feedback terms related to epidemics and epidemic control strengthened the base query thus reducing the ambiguity of the translation.

The results in Table 6.20 show that Spanish and English cross-language queries improve by up to 34% and 16%, respectively. In both cases, pre-translation feedback modification primarily improves precision. The best results for translations to English resulted with the addition of twenty feedback terms from the top ten documents, while the translation to Spanish were most improved by the additon of five terms from the top thirty documents. The English base queries can be expanded with more expansion terms than the Spanish base queries because they are longer and thus provide a less ambiguous base for expansion.The most relevant documents tend to be those containing the most query terms. Given that the TREC-3 queries are longer and may provide a stronger base for translation, a retrieval should produce more relevant documents at the top of the ranking than for the shorter TREC-4 queries. This should mean that the TREC-3 queries will have a better pool of expansion terms to choose from.

The next group of experiments investigate the effectiveness of query expansion prior to automatic translation via LCA. The results are compared to previous results using local feedback. Individual terms are translated word-by-word (WBW) and multi-term concepts are translated as phrases. In the event that no phrasal translation

	Translations to English			Translations to Spanish			
Feedback Terms	0	20	10	0	5	5	5
Training Documents	0	10	10	0	10	30	50
Average Precision:	0.0922	0.1072	0.0961	0.0823	0.1014	0.1099	0.1021
% Change:		16.4	4.3		23.2	33.5	24.0
Precision:							
5 docs:	0.2100	0.2300	0.2600	0.2000	0.2600	0.2500	0.2600
10 docs:	0.2050	0.2250	0.2300	0.2100	0.2500	0.2300	0.2600
15 docs:	0.2000	0.2233	0.2167	0.1867	0.2433	0.2400	0.2433
20 docs:	0.1900	0.2050	0.2075	0.1975	0.2300	0.2375	0.2350
30 docs:	0.1717	0.2050	0.1950	0.1900	0.2017	0.2217	0.2217

Table 6.20. Best pre-translation feedback results for TREC-3 and TREC-4 queries. (TREC-3 data consists of the TREC-3 English corpus, queries 151-162,164-171, and Spanish->English translations. TREC-4 data consists of the TREC-4 ISM Spanish corpus, queries 26-45, and English->Spanish translation.)

is found, phrases are translated WBW. Table 6.21 shows 4 representations of TREC-4 query SP29. First is the original query, second is the manual translation (base) including automatically identified phrases, third is the LCA expanded query, and fourth is the automatic translation of the third. Parentheses surround LCA expansion phrases and phrases automatically identified in the base query. Brackets surround the translation of each term or phrase.

Original Trec Query	las relaciones económicas y comerciales entre México y Canadá
Manual Translation to English	the economic and (commercial relations) between mexico and canada
Pre-translation Expanded Query	economic (commercial relations) mexico canada mexico (trade agreement) (trade zone) cuba salinas
MRD Translation of Pre-translation Expansion Query	[económico equitativo][comercio negocio tráfico industria] [narración relato relación][Méjico México] Canadá [Méjico México] [convenio comercial] [comercio negocio tráfico industria] zona cuba salinas

Table 6.21. Four query representations: original, base (with phrases in parentheses), LCA expanded base query (expansion terms in second row), WBW + phrasal translation of LCA expanded base.

First, we look at the effects of LCA expansion without phrasal recognition in the base query and compare a straight WBW translation of all concepts with a combination of phrasal and WBW translation. We then combine phrasal recognition in base queries with LCA expansion followed by both WBW and phrasal translation.

Translations of multi-term LCA pre-translation expansion concepts are wrapped in the INQUERY #passage25 and #phrase operators. For example, #passage25(#phrase(North American Free Trade Agreement)). Terms within a #phrase operator are evaluated to see whether they co-occur frequently in the collection. If they do, co-occurrences within 3 terms of each other are considered when calculating belief. If not, the terms are treated as having equal influence on the final result in order to allow for the possibility that individual occurrences are evidence of relevance. The #passage25 operator looks for the elements to occur within a window of 25. This operator ensures that terms which do not co-occur frequently be found a limited distance apart. The effectiveness of phrasal translations is highly sensitive to the quality of the translation. Although our phrase dictionary is based on phrasal information, it is also based on word usage information which may not produce quality phrasal translations. For this reason, we chose these operators to limit the effects of lower quality expansion-term translations.

The best results for automatic translations to Spanish are shown in Table 6.22. Descriptions of query processing for rows 2-7 follow. Row 2 (MRD) is the automatic word-by-word translation of the base (manual translation of original TREC-4) queries. For row 3, phrases were identified in the base queries and then WBW translation was augmented by phrasal translation (MRD + Phr). Row 4 shows results for pre-translation LCA expanded base queries that were translated word-by-word (MRD + LCA-WBW). Row 5 represents pre-translation LCA expanded base queries translated word-by-word with phrasal translation where possible (MRD + LCA-Phr). In Row 6, after phrase identification in base queries, they were expanded via LCA prior to translation. The expanded queries were then translated word-by-word with phrasal translation where possible. Finally, row 7 shows results for pre-translation local feedback expanded base queries after word-by-word translation (LF).

Method	Avg	%Change
MRD	0.0823	
MRD+Phr	0.0826	0.3
MRD+LCA-WBW	0.0969	17.7
MRD+LCA-phr	0.1009	22.7
MRD+Phr+LCA-phr	0.1053	27.9
LF	0.1099	33.5

Table 6.22. Average precision for pre-translation expansion results. (TREC-4 ISM Spanish corpus, queries 26-45, and English->Spanish translation)

The best results were gained after adding the top 30 concepts from the top 20 documents. They show that LCA expansion is effective, but WBW translation of LCA concepts yields only a 17% increase. This is probably due to the ambiguity introduced through the loss of multi-term concepts. Further improvements are given when phrases are identified in the base queries and when multi-term concepts are translated as phrases. If multi-term concepts are translated as phrases, effectiveness goes up by 5%. The addition of phrasal recognition in the base queries boosts cross-language effectiveness by an additional 5%. These results show that the use of phrasal translation can indeed improve effectiveness.

Pre-translation LCA expansion results are still not as good as those for pre-translation local feedback. This is surprising since comparisons of local feedback and LCA in the monolingual environment [134] have shown LCA to be more robust for query expansion.

We hypothesized that although most phrases added by LCA appear to be good phrases, they may lose their effectiveness when taken as individual terms. This happens when a phrasal translation fails and we are forced to translate the phrase word-by-word. In addition, poor phrases will also tend to have negative effects when translated word-by-word. Results in section 6.1.2 show that query effectiveness is highly sensitive to the accuracy of phrasal translation. To test the above hypothesis, we performed LCA expansion returning only the best single-term concepts. Expansion by individual terms removes the need to translate multi-term LCA concepts, so eliminates the negative effects of translating them poorly.

We found that in some cases, our hypothesis is supported. However, it is not consistent since multi-term expansion is sometimes more effective than single-term expansion. Table 6.23 gives a few examples of LCA expansion with single- and multi-term concepts compared to expansion with only single-term concepts. In this table, each of the expansions was done using the top 20 passages and the top 5 or 30 concepts. Automatic translation is given as a baseline. We believe the inconsistency is related to the types of multi-term concepts that are included in the expansion and on translation accuracy.

Method	Avg prec	%Change
MRD	0.0823	
LCA5-Phrasal	0.0819	-0.5
LCA5-Single	0.1051	27.7
LCA30-Phrasal	0.1053	27.9
LCA30-Single	0.1010	22.7

Table 6.23. Average precision for multi-term and single-term concept expansion. (TREC-4 ISM Spanish corpus, queries 26-45, and English->Spanish translation)

Table 6.24 shows the best pre-translation results for expansion via local feedback and for single-term expansion via LCA. This shows that LCA can be more effective than local feedback when used prior to translation; however, the selection of expansion concepts is critical.

6.2.2.2 Post-translation Expansion

Post-translation feedback was expected to be more effective than pre-translation feedback for the TREC-4 (English) cross-language queries. This is because statistics

	MRD	LF	LCA10-Single
Avg prec	0.0823	0.1099	0.1139
% Change:		33.5	38.5
Precision:			
5 docs:	0.2000	0.2500	0.3100
10 docs:	0.2100	0.2300	0.2750
15 docs:	0.1867	0.2400	0.2600
20 docs:	0.1975	0.2375	0.2350

Table 6.24. Best pre-translation local feedback and single-term LCA expansion results. (TREC-4 ISM Spanish corpus, queries 26-45, and English->Spanish translation.)

collected for their automatic translations (see Table 6.3) suggest that they contain many erroneous terms. Post-translation feedback should add more good terms which would help to reduce the effect of inappropriate translation terms.

TREC-4 query 28 did not show improvement when feedback terms were added prior to translation. However, feedback after translation improved effectiveness by 47% over MRD translation alone. The improvement is due in part to the inclusion of several terms related to commerce and helps to reduce the effects of ambiguity by de-emphasizing "outliers" or poor translations such as "porcelana". Table 6.25 shows the differences between four representations of query 28; all but the third are stemmed. Words grouped in parentheses are translation equivalents replacing one query word. Words in square brackets are post-translation expansion terms.

Experimental results are given in Table 6.26. Post-translation feedback modification tends to improve recall with English and Spanish cross-language queries showing improvements of up to 47.5% and 14.3%, respectively.

We suspect that the difference in effectiveness of the pre-translation and post-translation methods is primarily due to query length. The shorter TREC-4 queries are probably less well specified than the longer TREC-3 queries. This suggests that the latter would benefit more from the addition of good pre-translation expansion terms to create a stronger base query for translation. In addition, the translations of

Original stemmed TREC query	relacion econom comerc mex pais asiat japon chin cor
MRD translation from English	(econom equit rentabl)comerc(narr relat rel)(mej mex)asiat(pai patri camp region tierr)japon chin cor
MRD translation to Spanish of English pre-translation expansion query	(económico equitativo rentable)comercial(narración relato relación)mexico(laca japonesa)(porcelana loza)korea korean nuclear(sur mediodía)norte seoul soviet asia pyongyang japanese comunista(comercio negocio tráfico industria)asian (union enlace sindicato gremio obrero unión manguito unión) diplomático beijing unido península roh
post-translation feedback expansion of MRD-translated Spanish query	econom equit rentabl comerc narr relat rel mej mex asiat pai patri camp region tierr japon chin cor [pais export asi comerci singapur kong merc taiw hong product japones industr invers canada millon dol malasi estadounidens tailandi import]

Table 6.25. Five stemmed query representations for SP28: original TREC Spanish, MRD-translation of manual English translation back to Spanish, MRD-translation to Spanish after pre-translation feedback (unstemmed), post-translation feedback expanded Spanish query.

the TREC-3 queries are longer and contain a higher percentage of erroneous terms. This suggests that post-translation expansion would be more effective at reducing the effect of poor translations for TREC-3 queries.

In the following experiments, post-translation LCA expansion including both single and multi-term concepts was performed. When post-translation LCA expansion is performed, multi-term concepts are wrapped in INQUERY #passage25#phrase operators. The top ranked concept was added to a query with a weight of 1.0. Each additional concept was down-weighted by 1/100 with respect to the weight given its predecessor. This weighting scheme was shown to be effective in LCA experiments for the TREC-5 [128] evaluations. Table 6.27 shows the best results for post-translation expansion via local feedback and LCA. In this table, local feedback expansion was done by addition of the top 20 terms from the top 50 documents. LCA expansion was done by addition of the top 100 concepts from the top 20 passages. Table 6.28 shows 2 representations of one of these queries. The first is the base query and the second

	Translations to English				Translations to Spanish			
Feedback Terms	0	5	5	5	0	20	20	30
Training Documents	0	10	30	50	0	10	50	10
Average Precision:	0.0922	0.1252	0.1346	0.1359	0.0823	0.0910	0.0916	0.0913
% Change:		35.8	46.1	47.5		10.6	11.3	10.9
Precision:								
5 docs:	0.2100	0.2600	0.2400	0.2400	0.2000	0.2500	0.1800	0.2300
10 docs:	0.2050	0.2300	0.2300	0.2350	0.2100	0.1950	0.1850	0.1800
15 docs:	0.2000	0.1967	0.2267	0.2300	0.1867	0.1900	0.1800	0.1800
20 docs:	0.1900	0.1875	0.2125	0.2200	0.1975	0.1975	0.1575	0.1800
30 docs:	0.1717	0.1750	0.1950	0.2000	0.1900	0.1633	0.1483	0.1617

Table 6.26. Best post-translation feedback results. (Data consists of the TREC-3 English corpus, queries 151-162,164-171, with Spanish->English translations and the TREC-4 ISM Spanish corpus, queries 26-45, with English->Spanish translations.)

the automatic translation of the base query. The last row gives the top 20 expansion concepts that were added to this query, with multi-term concepts in parentheses. Note that all terms are stemmed.

	MRD	LF	LCA20
Avg prec	0.0823	0.0916	0.1022
% Change:		11.3	24.1
Precision:			
5docs:	0.2000	0.1800	0.2200
10 docs:	0.2100	0.1850	0.2100
15 docs:	0.1867	0.1800	0.2167
20 docs:	0.1975	0.1575	0.2050

Table 6.27. Best post-translation local feedback and LCA expansion results. (Data consists of the TREC-4 ISM Spanish corpus, queries 26-45, with English->Spanish translations.)

The best post-translation LCA expansion is 11.6% more effective than the best post-translation local feedback expansion. Eleven of 20 queries do better with LCA as compared to 7 which do better with LF. A paired sign test shows this difference to be significant at p = .01. This supports earlier work by Xu [134] which showed LCA to be a more effective query expansion technique than local feedback. LCA is not always more effective than local feedback when pre-translation expansion is performed. This difference between effectiveness for pre-translation versus post-translation application

Manual Translation to English	economic commercial relations mexico european countries
MRD Translation to Spanish	comerc narr relat rel econom equit rentabl pai patri camp region tierr mej mex europ
Top 20 Post-translation Expansion Concepts	(est un) canada pai europ franci (diversific comerc) mex polit pais alemani rentabl oportun product apoy australi (merc europ) agricultor bancarrot region (comun econom europ)

Table 6.28. Two query representations for TREC query SP26: BASE and MRD translation of BASE. Row 3 gives the top 20 post-translation LCA expansion concepts for this query.

of the approaches is probably due to the confounding effects of translation ambiguity when pre-translation expansion is performed.

6.2.2.3 Combined Expansion

In the previous two sections, we show that pre- and post-translation expansion are effective means for reducing the negative impact of translation ambiguity. We expect that combining pre- and post-translation expansion will be more effective than either approach alone. In this section, we combine pre- and post-translation in the following way. First base queries are modified via feedback. Second, the modified queries are translated via MRD to the target language. Finally, the resulting queries from step two are modified via post-translation feedback and a retrieval evaluation is performed.

The combined method was most effective on the TREC-4 cross-language queries yielding up to a 51% improvement in average precision. The query sets for other combined-feedback runs show similar results. As would be expected, both precision and recall are improved by the combined method. The improvements occur because better query terms are added after the final feedback. Those terms tend to fine tune the query and de-emphasize inappropriate definitions. TREC-3 queries showed more than 40% improvement after combined feedback. Results are given in Table 6.29.

	Translations to Spanish			Translations to English			
Feedback Terms	0	10	20	0	5	20	5
Training Documents	0	30	50	0	30	30	20
Average Precision:	0.0823	0.1166	0.1242	0.0922	0.1372	0.1375	0.1366
% Change:		41.7	51.0		48.8	49.2	48.2
Precision:							
5 docs:	0.2000	0.2400	0.2600	0.2100	0.2600	0.2700	0.2400
10 docs:	0.2100	0.2200	0.2200	0.2050	0.2450	0.2500	0.2500
15 docs:	0.1867	0.2200	0.2000	0.2000	0.2433	0.2467	0.2400
20 docs:	0.1975	0.2075	0.2125	0.1900	0.2400	0.2350	0.2375

Table 6.29. Best combined pre-translation and post-translation feedback results. (Data consists of the TREC-4 ISM Spanish corpus, queries 26-45, with English->Spanish translations and the TREC-3 English corpus, queries 151-162,164-171, with Spanish->English translations.)

For both the Spanish and English base queries, those queries that showed improvement via pre-translation alone gained greater improvements from subsequent post-translation feedback. This suggests that the pre-translation feedback stage creates a better base for translation and then the post-translation stage reduces the negative effects of ambiguity caused by inappropriate term definitions.

The following combination experiments start with the pre-translation LCA expansion of the base queries. After the expanded queries are translated automatically, they are expanded again via LCA multi-term expansion. The base query set for the post-translation expansion phase in these experiments is the best pre-translation, single-term concept LCA expanded query set, as described in Section 6.2.2.1. Table 6.30 shows four representations of one of these queries. First is the original query, second is the manual translation (BASE) including automatically identified phrases, third is the pre-translation LCA single-term expanded query, and fourth is the automatic translation of the third. The last row gives the top 20 post-translation expansion concepts that were added to this query, with multi-term concepts in parentheses. Note that all terms are stemmed. Parentheses surround LCA expansion phrases and phrases automatically identified in the BASE query. Brackets surround the translation of each term or phrase.

Original TREC Query	las relaciones económicas y comerciales entre México y Canadá
Manual Translation to English	the economic and (commercial relations) between mexico and canada
Pre-translation Expanded Query	economic (commercial relations) mexico canada mexico free-trade canada trade mexican salinas cuba pact economies barriers
MRD Translation of Pre-translation Expanded Query	[económico equitativo][comercio negocio tráfico industria] [narración relato relación] [Méjico México] Canadá [Méjico México][convenio comercial] [comercio negocio tráfico industria] zona cuba salinas
Top 20 Post-translation Expansion Concepts	canada (libr comerci) trat ottaw dosm (acuerd paralel) norteamer (est un) (tres pais) import eu (vit econom) comerci (centr econom) (barrer comerc) (increment subit) superpot rel acuerd negoci

Table 6.30. Four query representations: original, base (with identified phrases), LCA expanded base, WBW + phrasal translation of LCA expanded base.

The combined LCA approach is more effective than either pre- or post-translation LCA expansion alone. This was also shown to be the case for local feedback expansion. Table 6.31 gives results for automatic translation, the best combined pre- and post-translation local feedback expansion, and the best combined LCA expansion. In this experiment, queries were expanded by the top 50 terms from the top 20 passages in the post-translation LCA phase. Fourteen and eleven queries show improvement over MRD translation alone for LCA and LF, respectively. The LCA approach shows a 9% greater improvement than the local feedback approach, but this difference is not statistically significant. When the two methods are compared, nine queries do better with LCA expansion as compared to 10 that do better with LF expansion. However, it is interesting to compare the effects of LCA and local feedback expansion on precision. The LCA expansion has higher precision at low recall levels. This is important in a CLIR environment. A person may not be proficient at reading a foreign language, so could not be expected to look through more than the top retrieved documents.

	MRD	LF	LCA20-50
Avg prec	0.0823	0.1242	0.1358
% Change:		51.0	65.0
Precision:			
5 docs:	0.2000	0.2600	0.3700
10 docs:	0.2100	0.2200	0.2850
15 docs:	0.1867	0.2000	0.2767
20 docs:	0.1975	0.2125	0.2600

Table 6.31. Best combined pre- and post-translation local feedback and LCA expansion results. (Data consists of the TREC-4 ISM Spanish corpus, queries 26-45, with English->Spanish translations.)

6.2.2.4 Parallel Corpus Disambiguation

Parallel corpora contain a set of documents and their translations into one or more other languages. Analysis of these paired documents can be used to infer the most likely translations of terms between languages in the corpus. We employ parallel corpus analysis to look at the impact of query term disambiguation on CLIR effectiveness. The technique is a modification of one used by NMSU [38] and is described below.

Source language (Spanish) queries are first tagged using a part-of-speech (POS) tagger. Each Spanish source term is replaced by all possible target language (English) translations for the term's POS. If there is no translation corresponding to a particular query term's tag, the translations for all parts-of-speech listed in the dictionary for that term are returned. There may be one or more ways to translate a given term. When more than one equivalent is returned, the best single term is chosen via parallel corpus disambiguation.

Disambiguation proceeds in the following way. The top 30 Spanish documents are retrieved from the parallel UN corpus in response to a Spanish query. The top 5000 terms based on Rocchio ranking are extracted from the English UN documents that correspond to the top 30 Spanish documents. The translations of a query term are ranked by their score in the list of 5000. The highest ranking translation(s) is chosen as the “best“ translation for that term. If none of the equivalents are on the list, no disambiguation is performed and all equivalents are chosen. This method differs from that of NMSU in two ways. First, we used document level alignment instead of sentence level alignment. Second, rather than disambiguation based on the top documents retrieved in response to the query, they retrieved the top sentences in response to a query term. They then chose the term translation that retrieved the most sentences like those retrieved for the untranslated term.

6.2.2.5 Disambiguating Phrases via Unaligned Corpora

The correct translations of query terms should co-occur in target language documents and incorrect translations should tend not to co-occur. We use this hypothesis as the foundation for a method to disambiguate phrase translations. Given the possible target equivalents for source phrase terms, we infer the most likely translations by looking at the pattern of co-occurrence for each possible combination of definitions. Our method is similar to that of Dagan described in Section 5.3.3. A description of our method follows. The example describes how to disambiguate a two word phrase, but the approach can be applied to any number of terms.

Given two tagged source terms, collect all target translation equivalents for each term. Generate all possible sets $\{a, b\}$ such that a is a definition of $term1$ and b is a definition of $term2$. Measure the importance of co-occurrence of the elements in a set by the *em* metric [135]. It is a variation of EMIM [126] (described in Section 1.2.3.3) and measures the percentage of the occurrences of a and b that are net co-occurrences (co-occurrences minus expected co-occurrences), but unlike EMIM does not favor uncommon co-occurrences.

$$em(a, b) = max(\frac{n_{ab} - En(a, b)}{n_a + n_b}, 0)$$

where n_a, n_b are the number of occurrences of a and b in the corpus, and n_{ab} is the number of times both a and b fall within a text window of t words. $En(a,b) = \frac{n_a n_b}{N}$ and N is the number of text windows in the corpus. Each set is ranked by em score and the highest ranking set is taken as the appropriate translation. If more than one set has a rank of one, all of them are taken as translations. Our method differs from that of Dagan, et al. in the following ways. They paired words to be translated via syntactic relationships e.g. subject-verb after parsing. Selection was made via a statistical model based on the ratio of the frequency of co-occurrence of one alternative to the frequency of co-occurrence for all the alternatives. Our algorithm is given below. Let $t =$ the number of terms in the phrase to be translated, $position_t =$ the list of positions that term t can be found in, $win\_size =$ the size of the window within which translations must be found, $win\_limit =$ the end of current window, $win\_count =$ the number of windows within which this translation is found, and $i = 1$ (first term in phrase).

```
$win\_limit = win\_size$;
for each $position(i)$ {
    if $position(i) > win\_limit$
        $win\_limit$ += $win\_size$;
    for (n = 2; n <= t; n++) {
        if position(n) < $win\_limit$
            check position of next phrase term;
        else if (no location of n is in window)
            break;
    }
    if (all terms found in window ){
        $win\_cnt + +$;
        $win\_limit$ += $win\_size$;
    }
}
```

The complexity of this algorithm is $O(nl(log\ n))$ where l is the number of terms in the phrase to be translated and n is the length of the largest number of co-occurrences for any term. The outer loop begins by iterating across the positions of the first term of a possible phrasal translation. This outer loop looks to identify potential occurrences of the phrase terms within the term window. With each pass, the inner loop tries to add another term to the potential match. If the end of a position list is reached, no more matches are possible. If the inner loop ends because it can't complete a match, the outer loop begins the next potential match. Once a match has been found in a window, the search for matches continues in the next window.

As mentioned above, translating multi-term concepts as phrases is an important step in reducing translation error. In these experiments, we compare the ability of our phrase dictionary with that of the co-occurrence method (CO) (as described in the previous section) to translate phrases. We use co-occurrence statistics to reduce ambiguity by inferring the correct translation of phrases not translatable via our phrase dictionary and compare the effectiveness of the two methods with word-by-word translation as a baseline.

Given the phrases in the CSE query set, we compared the number for which translations could be found in the phrase dictionary with those translatable via CO. The comparison was done by a human assessor who determined whether phrasal translations via either method were correct. Thirty-three phrases were identified in seventeen out of twenty-one TREC-6 queries. Ten phrases were duplicates leaving only twenty-three unique phrases. Table 6.32 shows statistics for the types of phrases identified and also gives results of the comparison. The first row shows the number and types of phrases. The second and third rows show the numbers of phrases of each type that are translatable via our phrase dictionary and co-occurrence method respectively.

	Unique	Compositional	Non-compositional
	23	21	2
Phr. Dict	8	6	2
Co-occur.	13	13	N/A

Table 6.32. Breakdown of total number of phrases and phrase types in short Spanish queries, including the numbers translatable via phrase dictionary or co-occurrence method. (TREC-6 Spanish queries 1,2,4-7, 9-14, 16-24)

Translations of phrases found in the phrase dictionary are correct. Note that the six compositional phrases found in the phrase dictionary can also be correctly translated via CO. CO will only work for the translation of compositional phrases. For example, the Spanish phrase *medio oriente* is compositional as it can be translated word-by-word as *middle east.* However,the phrase *contaminación del aire* can not be translated compositionally into *air pollution* since *pollution* is not a translation of *contaminación.* Therefore, we rely upon our phrase dictionary for the translation of non-compositional phrases.

Thirteen compositional phrases are translated correctly using the co-occurrence method. For example, *abuso infantil, comercio marfil, proceso paz* are correctly translated to *child abuse, ivory trade, and peace process*, respectively. The possible translation sets for *processo paz* can be generated from the translations of the constituent terms. The target equivalents of *proceso* and *paz* are *process, lapse of time, trial, prosecution, action, lawsuit, proceedings, processing* and *peace, peacefulness, tranquility, peace, peace treaty, kiss of peace, sign of peace*, respectively. The translation of one of the thirteen is not ambiguous since both constituent source terms have only one target translation.

Seven other compositional phrases were not in the phrase dictionary and were translated incorrectly via CO. In these cases, the translation failure does not appear to be a big problem since only one of the queries containing a poorly translated phrase loses effectiveness. This may be due to the following. First, some of the poorly translated phrases are not very important to the queries that contain them. *Mejor artículo* means *best item*, but is translated as *best thing*. Second, at least one of the constituent term translations for each poorly translated phrase is correct. The effect of disambiguating at least one of the terms may reduce the overall negative effect of failing to translate the phrase. The phrase *prueba de inflación* meaning *inflation proof* was translated as *inflation evidence*. In this case, the key term *inflación* was translated correctly. Table 6.33 gives the effect that translating phrases had on effectiveness of MSE Spanish queries . It shows precision values for word-by-word with phrase dictionary translation (PD) versus word-by-word with co-occurrence translation (CO) and word-by-word with phrase dictionary and co-occurrence translation (PD+CO) as compared to the baseline of word-by-word (WBW) translation. Each of the queries containing correct CO phrasal translations improved. The improvement in effectiveness with the addition of CO over PD alone is significant at the .01 level. The addition of phrasal translations using both methods brings cross-language effectiveness up to 79% of monolingual as measured by average precision. In fact, only half of the queries in which phrases were translated via co-occurrence information do worse than their monolingual counterparts. Translation without phrases yields only 60% of monolingual.

It should be noted that poor translations can decrease effectiveness as shown in [13]. This is the case for some of the phrasal translations in the MSE queries, where translation were made from Spanish to English. Phrasal translations are moderately effective, although the improvement is not statistically significant. Tables 6.34 and 6.35 show recall-precision values and phrase statistics for the MSE query set. In some cases although the translation is correct, replacement of the word-by-word translation

Query	WBW	PD	CO	PD+PLC	PD+CO
Avg.Prec.	0.2331	.2944	0.2741	0.2551	0.3057
% change		26.3	17.6	9.4	31.1
Precision at:					
5 docs:	0.3619	0.3905	0.3714	0.4095	0.4190
10 docs:	0.3286	0.3714	0.3762	0.3857	0.4048
20 docs:	0.3095	0.3738	0.3690	0.3524	0.4048
30 docs:	0.2810	0.3413	0.3238	0.3254	0.3746

Table 6.33. Average precision for word-by-word translations and word-by-word translations augmented by both phrasal translation methods. (TREC-6 AP English corpus, Spanish queries 1,2,4-7, 9-14, 16-24, and Spanish-> English translations.)

with the phrasal translation reduces effectiveness. A query by query analysis explains why.

Query	WBW	PD	PD+CO
Avg.Prec.	0.1529	0.1636	0.1583
% Change:		7.0	3.5
Precision at:			
5 docs:	0.4300	0.4100	0.3600
10 docs:	0.3950	0.3850	0.3450
20 docs:	0.3400	0.3350	0.3100
30 docs:	0.3050	0.2983	0.2917

Table 6.34. Average precision for word-by-word translations and word-by-word translations augmented by both phrasal translation methods. (TREC-3 English corpus, queries 151-162,164-171, and Spanish->English translations.)

	Unique	Compositional	Non-compositional
	41	29	12
Phr. Dict	8	3	5
Co-occur.	21	21	N/A

Table 6.35. Breakdown of total number of phrases and phrase types in MSE Spanish queries, including the numbers translated correctly via phrase dictionary or co-occurrence method. (TREC-3 Spanish queries 151-162,164-171.)

The effectiveness is improved for eight queries and degraded for four queries by co-occurrence translation in the MSE query set. For two of the failures, the translation is incorrect and results in one of the important concepts of the query being lost. A third query is degraded because the manual translation was incorrect. The original query was about generic drugs. When manually translated to Spanish, generic was translated as “marca libre” rather than “generico”. The fourth query loses effectiveness despite the fact that the translation is correct. This is because the word-by-word translation of the query gave an expansion effect. Co-occurrence suggested “coche eléctrico” be translated to “electric car” which is correct. However, the word-by-word translation of “coche” contains: car, motorcar, and automobile. Weighting schemes for giving credit to both phrasal translations and word-by-word translations may be a way to avoid this pitfall. It may also be possible to take a range of *em* values rather than an absolute maximum that will reduce ambiguity by eliminating erroneous translations, but still allowing the expansion effect by choosing more translations when applicable. However, despite the fact that co-occurrence translation of phrases may result in a slight drop in effectiveness as compared to that of phrase dictionary translations alone, in Section 6.2.3, in Table 6.40 we show that this is not the case when other methods of ambiguity reduction are also applied. Although average precision may be slightly lower for queries employing co-occurrence translation of phrases, they provide a much better base query for further ambiguity reduction.

One way to reduce the problem of incorrect translations could be to include more query terms in the co-occurrence analysis. Including more terms would provide more context and may further disambiguate translations. In particular, the inclusion of additional terms having unambiguous translations themselves would provide “anchor points”. These anchor point would help to establish the correct context for the disambiguation.

6.2.2.6 Comparing CO and PLC Methods for Term Disambiguation

Parallel corpora can be used to disambiguate term translations as described in section 6.2.2.4. We showed in the above section that co-occurrence statistics can be used to disambiguate terms as phrasal constituents. We now show that it can also be used for general term disambiguation and compare it to the parallel corpus technique.

We translated our query set in the following way. Phrases were translated using the phrase dictionary. Terms were translated word-by-word and then disambiguated using the parallel corpus method. We looked at sixty terms disambiguated by the parallel corpus and investigated how well they could be disambiguated via co-occurrence. We used the same co-occurrence method that was used for disambiguating phrase translations. However, rather than require the term to be a phrase constituent, we paired the term to be disambiguated with an anchor. In this investigation, an anchor is chosen randomly and is a query noun that has an unambiguous translation, a proper noun, or a phrase translation. The resulting translations were then evaluated by a human assessor. Our conjecture was that co-occurrence disambiguation would not do any worse than parallel corpus disambiguation. Table 6.36 shows the overlap of terms correctly and incorrectly disambiguated by each method.

	correctly disamb. via parallel corpus	incorrectly disamb. via parallel corpus	Total
correctly disamb. via co-occurrence	36	11	47
incorrectly disamb. via co-occurrence	3	10	13
Totals	39	21	60

Table 6.36. Term disambiguation results for 60 word pairs. (Word pairs were extracted from TREC-6 Spanish queries 1,2,4-7, 9-14, 16-24.)

Let PC be the number of words which were correctly disambiguated via PLC and incorrectly disambiguated via CO and let CP be the reverse. The null hypothesis is that PC pairs are as likely as CP pairs. A McNemar's test shows that we can reject the null hypothesis with a significance level of $p \leq 0.057$. When the co-occurrence method does not correctly disambiguate a term, there appears to not be enough context to infer the correct translation. The translation of *Efectos del chocolate en la salud. Cuales, si existen, son los efectos del chocolate en la salud.* is *The effects of chocolate on health. What, if any, are the effects of chocolate on health?.* The Spanish word *chocolate* can be translated as *chocolate, cocoa,* or *blood.* Given that it is more common to find *blood* co-occurring with *health*, blood is chosen over the uncommon and correct translation *chocolate.* One means of alleviating the problem could be through pre-translation expansion. This is described in more detail later, but the basic idea follows. Prior to translation, retrieval is performed with the source query on a source language database. The query is then expanded with the best terms from the top ranking passages retrieved in response to the query. These expansion terms may provide enough context to be good anchors for disambiguation. *Hershey*, a brand of chocolate, is one of the expansion terms for the example query given above. Using *Hershey* as an anchor, rather than one of the original query terms, will more likely disambiguate Sp. *chocolate* to *chocolate* than to *blood.* This is because *Hershey* is more likely to be found in contexts related to chocolate than in contexts related to blood.

The failure of the parallel corpus method to disambiguate seems to be related to there being few or no documents related to the query. This is a problem more likely to happen the narrower or the more different the domain of the parallel corpus is from the corpus being searched. Our experiments are based on the UN parallel corpus which contains documents concerned with international peace and security, and health and education in developing countries. The query set is more general. Although there will be some general vocabulary overlap, the lack of relevant documents may prevent the disambiguation of query specific concepts. The UN corpus does not, for example, contain any documents relating to the effects of chocolate on health and the parallel corpus method incorrectly disambiguates *chocolate* to *blood.* Of course this remains conjecture and needs to be tested experimentally. However, it suggests that the co-occurrence method will be a more effective disambiguation method than the parallel corpus technique. This may be especially true when we cannot rely on domain specific resources or at least on there being more domain overlap.

Nearly all of the phrases not translatable via the phrase dictionary are translatable word-by-word. We were interested in comparing the effectiveness of parallel corpus disambiguation with co-occurrence disambiguation. Recall that for all queries, terms are translated word-by-word and noun phrases are translated via our phrase dictionary. The co-occurrence method (CO) disambiguates the remaining phrase term translations based on their co-occurrence with other terms in a phrase. The parallel corpus disambiguation method (PLC) uses query context to disambiguate all remaining terms whether or not they are constituents of a phrase. We also wanted to see how the PLC and CO methods compared to more sophisticated machine translation (MT) systems.

Using a baseline of word-by-word translation (WBW), Table 6.37 compares the effectiveness of both PLC and CO with that of two MT systems. The first is a web accessible off-the-shelf package called T1 from Langenscheidt [55] and the second is the on-line SYSTRAN [119] system. This table also gives cross-language performance as a percentage of monolingual. The co-occurrence method is more effective and gives higher recall and higher precision at all recall levels than does the PLC method. The SYSTRAN MT system is about as effective as the PLC method. There is no significant difference between the Langenscheidt MT system and the CO method which attains 79% of monolingual effectiveness. This is encouraging because it shows that co-occurrence information can be successfully employed to attain the effectiveness of a reasonably effective MT system. This is a positive statement for the possibilities of cross-language searching in languages for which few resources exist or for which a reasonable MT system does not exist.

Method	Precision	%change	% Monolingual
Monolingual	0.3869		-
WBW	0.2331		60
PLC	0.2551	9.4	65
CO	0.3057	31.1	79
T1	0.3066	31.5	79
SYSTRAN	0.2584	10.8	67

Table 6.37. Average precision as a percentage of that for monolingual. (Data consists of the TREC-6 AP English corpus and Spanish queries 1,2,4-7, 9-14, 16-24, with Spanish->English translations.)

6.2.3 Combinations of Disambiguation Methods

In the following experiments, we look at the effectiveness of combining the disambiguation methods described above with query expansion via Local Context Analysis (LCA). We translated queries automatically employing both POS and SYN disambiguation as in section 6.2.1. Phrases were translated using the phrase dictionary and then one of the corpus disambiguation methods was applied. The co-occurrence method was performed with a window size of 250 terms. Queries were then expanded via LCA prior to translation, after translation or both before and after translation. We also compared these results to the expansion of queries translated via the "sense1" method for which results were shown in Sections 6.1.1, 6.2.2.1, 6.2.2.2, and 6.2.2.3. Recall that for the sense1 method, only the target translations corresponding to the first sense listed in the dictionary entry are taken. This is done in an effort to reduce the number of extraneous terms.

6.2.3.1 Pre-translation Expansion

The following set of experiments show how effective pre-translation expansion is for further disambiguating three types of query translations: the sense1 method, the parallel corpus disambiguation method (PLC), and the co-occurrence method (CO). Pre-translation expansion is done in the following way. The top 20 passages are retrieved in response to the source query. The top 5 source terms are then added to the query. Expansion is followed by query translation. Average precision values are given in table 6.38. Word-by-word translation as described in section 6.2.1 is used as a baseline. Columns two, four, and six are queries translated via the sense1, PLC, and CO methods, respectively. Columns three, five, and seven are the sense1, PLC, and CO methods each with pre-translation expansion. Earlier work showed that pre-translation expansion enhances precision. Results are consistent with this, with the exception of pre-translation expansion of the PLC disambiguated queries. The problem here is that many of the expansion terms were disambiguated incorrectly, so that nearly half of the queries lost effectiveness. The improvement in average precision of expanded co-occurrence disambiguated queries over co-occurrence disambiguation alone is not significant. This may be due to the improved quality of CO translation over the other translation methods. In other words, the CO method alone may be reducing much of the ambiguity that is reduced by pre-translation expansion with other methods of translation.

Query	WBW	1st	1st+Pre	PLC	PLC+Pre	Co	Co+Pre
Avg.Prec.	0.2331	0.2392	0.2568	0.2551	0.2155	0.3057	0.3098
% change		2.6	10.1	9.4	-7.6	31.1	32.9
Precision at:							
5 docs:	0.3619	0.3238	0.3429	0.4095	0.3333	0.4190	0.4667
10 docs:	0.3286	0.2810	0.3190	0.3857	0.3476	0.4048	0.4333
20 docs:	0.3095	0.3119	0.2952	0.3524	0.3143	0.4048	0.3976
30 docs:	0.2810	0.2651	0.2714	0.3254	0.2857	0.3746	0.3683
100 docs:	0.1705	0.1676	0.1795	0.1929	0.1652	0.2443	0.2324

Table 6.38. Average precision and precision at low recall for word-by-word, sense1, sense1 with pre-translation expansion, parallel corpus disambiguation, parallel corpus disambiguation with pre-translation expansion, co-occurrence disambiguation, and co-occurrence disambiguation with pre-translation expansion. (Data consists of the TREC-6 AP English corpus and Spanish queries 1,2,4-7, 9-14, 16-24, with Spanish-> English translations.)

6.2.3.2 Post-translation Expansion

In these experiments, post-translation LCA expansion was performed by addition of the top 50 concepts from the top 30 passages. Expansion was carried out after query translation via either the sense1, PLC, or CO methods.

Table 6.39 shows average precision values for seven query sets. As in the previous section, word-by-word translation is used as a baseline. Columns three, five, and seven are the sense1, PLC, and CO methods, each with post-translation expansion. Our earlier work showed that post-translation expansion enhances recall and precision. These results are consistent with those findings. The most effective queries are those translated via CO followed by post-translation expansion. Recall is also higher for this query set.

Query	WBW	1st	1st+Post	PLC	PLC+Post	Co	Co+Post
Avg.Prec.	0.2331	0.2392	0.3317	0.2551	0.2864	0.3057	0.3623
% change		2.6	42.3	9.4	22.8	31.1	55.4
Precision at:							
5 docs:	0.3619	0.3238	0.4476	0.4095	0.4000	0.4190	0.4857
10 docs:	0.3286	0.2810	0.4333	0.3857	0.3857	0.4048	0.4857
20 docs:	0.3095	0.3119	0.3905	0.3524	0.3667	0.4048	0.4429
30 docs:	0.2810	0.2651	0.3651	0.3254	0.3476	0.3746	0.4111
100 docs:	0.1705	0.1676	0.2452	0.1929	0.2167	0.2443	0.2838

Table 6.39. Average precision and precision at low recall for word-by-word, sense1, sense1 with post-translation expansion, parallel corpus disambiguation, parallel corpus disambiguation with post-translation expansion, co-occurrence disambiguation, and co-occurrence disambiguation with post-translation expansion. (Data consists of the TREC-6 AP English corpus and Spanish queries 1,2,4-7, 9-14, 16-24, with Spanish-> English translations.)

Recall from Section 6.2.2.5, results showed that effectiveness of phrasal translation alone was sometimes better than combining it with phrase translation disambiguation via co-occurrence. However we find that although co-occurrence may yield a moderate reduction in effectiveness, when combined with query expansion, effectiveness is significantly better when phrasal translations are generated via both co-occurrence and the phrase dictionary. Table 6.40 gives recall-precision figures to compare phrasal translation via phrase dictionary with and without co-occurrence disambiguation of phrases. It also shows effectiveness of both after further disambiguation via post-translation LCA expansion. The moderate reduction in effectiveness when co-occurrence is applied without expansion is due in part to the loss of expansion effect that occurs when query term translations contain more than one appropriate translation - for example, the translation of *coche* in Spanish to *car, automobile,* and *automobile.* However, the CO method yields a less ambiguous query and a better base for expansion than the query to which CO is not applied. This results in better expansion terms in the former and thus greater effectiveness. The result is significant at p $\leq$ 0.012.

Query	WBW+Phr	WBW+Phr+Post LCA30-50	WBW+Phr +Co	WBW+Phr+Co +Post LCA30-50
Relevant Docs:	3255	3255	3255	3255
Relevant Ret:	913	1435	939	1453
Avg Prec	0.1636	0.2152	0.1583	0.2367
% Change:		31.6	-3.3	44.7
Precision:				
5 docs:	0.4100	0.4200	0.3600	0.4400
10 docs:	0.3850	0.4050	0.3450	0.4450
15 docs:	0.3333	0.3933	0.3400	0.4400
30 docs:	0.2983	0.3467	0.2917	0.3783
100 docs:	0.1930	0.2370	0.1895	0.2465

Table 6.40. Average precision and precision at low recall for MSE queries with word-by-word translations augmented by the phrase dictionary, phrase dictionary plus LCA expansion, phrase dictionary plus co-occurrence disambiguation, and phrase dictionary plus co-occurrence disambiguation and LCA expansion. (Data consists of the TREC-3 English corpus and queries 151-162,164-171, with Spanish->English translations.)

6.2.3.3 Combined Pre- and Post-translation Expansion

The combination experiments start with the pre-translation LCA expansion of the source queries. After the expanded queries were translated automatically via the sense1, PLC, or CO method, they were expanded again via LCA multi-term expansion. The pre- and post- translation phrases proceed as described in Sections 6.2.3.1 and 6.2.3.2. Results are given in Table 6.41.

Query	WBW	1st	1st+Comb	PLC	PLC+Comb	Co	Co+Comb
Avg.Prec.	0.2331	0.2392	0.3193	0.2551	0.2593	0.3057	0.3533
% change		2.6	37.0	9.4	11.2	31.1	51.5
Precision at:							
5 docs:	0.3619	0.3238	0.3905	0.4095	0.3619	0.4190	0.4952
10 docs:	0.3286	0.2810	0.4190	0.3857	0.3333	0.4048	0.4810
20 docs:	0.3095	0.3119	0.4024	0.3524	0.3357	0.4048	0.4452
30 docs:	0.2810	0.2651	0.3556	0.3254	0.3095	0.3746	0.3968
100 docs:	0.1705	0.1676	0.2424	0.1929	0.2019	0.2443	0.2690

Table 6.41. Average precision and precision at low recall for word-by-word, sense1, sense1 with post-translation expansion, parallel corpus disambiguation, parallel corpus disambiguation with post-translation expansion, co-occurrence disambiguation, and co-occurrence disambiguation with post-translation expansion. (Data consists of the TREC-6 AP English corpus and Spanish queries 1,2,4-7, 9-14, 16-24, with Spanish-> English translations.)

As expected, combining pre- and post-translation expansion boosts both precision and recall. There is no significant difference between post-translation and combined expansion of the CO translated queries. This makes sense in light of the fact that the CO method appears to disambiguate queries so well that pre-translation expansion has little impact on effectiveness. There is no significant difference between CO expanded via the post-translation method or CO expanded via the combined method. However, the post-translation expansion method may be preferred here since precision is slightly higher at low recall.

Combining pre- and post-translation LCA expansion is also effective for the MSE queries as shown in Table 6.42. This result is significant at the level $p \leq 0.08$.

Table 6.43 gives recall-precision figures for the MSE queries after combining all of the disambiguation techniques and compares them to monolingual effectiveness with and without expansion. Combining word-by-word translation with phrase translation via dictionary and co-occurrence disambiguation, synonym operator, POS, and combined pre- and post-translation LCA expansion yields effectiveness that is 10% better than monolingual without expansion. When monolingual queries are augmented by

Query	No LCA	LCA
Relevant Docs:	3255	3255
Relevant Ret:	939	1565
Avg Prec	0.1583	0.2489
% Change:		57.3
Precision at:		
5 docs:	0.3600	0.4500
10 docs:	0.3450	0.4500
20 docs:	0.3100	0.4175
30 docs:	0.2917	0.3833
100 docs:	0.1895	0.2605

Table 6.42. Average precision and precision at low recall for word-by-word translations of MSE queries augmented by phrasal dictionary and co-occurrence phrase disambiguation with and without combined LCA Expansion. (Data consists of the TREC-3 English corpus and queries 151-162,164-171, with Spanish->English translations.)

LCA expansion, the cross-language queries still achieve 82% monolingual effectiveness.

Table 6.44 shows the effectiveness of each of the best expansion methods applied to the CSE queries as a percentage of monolingual performance as measured by average precision. Results show that combining our disambiguation methods brings cross-language performance to more than 90% of monolingual performance for these queries.

Query			
Relevant Docs:	3255	3255	3255
Relevant Ret:	1800	2145	1565
Avg Prec	0.2259	0.3035	0.2489
% Change:		34.4	10.2
Precision:			
5 docs:	0.4900	0.6300	0.4500
10 docs:	0.4650	0.5700	0.4500
20 docs:	0.4475	0.5300	0.4175
30 docs:	0.4250	0.4900	0.3833
100 docs:	0.2910	0.3365	0.2605

Table 6.43. Average precision and precision at low recall for long queries (monolingual, monolingual+expansion, and cross-language combined). Data consists of the TREC-3 English corpus and queries 151-162,164-171, with Spanish->English translations.

Method	Precision	% Monolingual
Mono	0.3869	-
sense1+post	0.3317	86
CO+pre	0.3098	80
CO+combined	0.3533	91
CO+post	0.3623	94

Table 6.44. Average precision as a percentage of that for monolingual. (Data consists of the TREC-6 AP English corpus and Spanish queries 1,2,4-7, 9-14, 16-24, with Spanish->English translations.)

6.3 Summary and Discussion

One of the main hurdles to improving cross-language retrieval effectiveness has been the reduction of ambiguity associated with query translation. Translation error is due largely to addition of extraneous terms and failure to correctly translate phrases. In addition, the resources needed to address this problem typically require considerable manual effort to construct and may be difficult to acquire. Automatic dictionary translations are attractive because they are cost effective and easy to perform, and resources are readily available.

A few simple techniques such as part-of-speech tagging and the use of the #synonym operator can address the extraneous term problem. Phrasal translation is more problematic. Certain types of multi-term concepts, such as proper noun phrases, are easily translated via MRD. However, dictionaries do not provide enough context for accurate phrasal translation in other cases. The correct translations of phrase terms tend to co-occur and incorrect translations tend not to co-occur. Corpus analysis can exploit this information to significantly reduce ambiguity of phrasal translations. Combining phrase translation via phrase dictionary and co-occurrence disambiguation brings CLIR performance up to 79% of monolingual. The co-occurrence technique can also be used to reduce ambiguity of term translations.

Query expansion is another means of resolving ambiguity. LCA expansion gives higher precision at low recall levels, which is important in a CLIR environment.

LCA, which typically expands queries with phrases in addition to single terms, is more sensitive to translation effects when pre-translation expansion is performed. This is because phrases that must be translated WBW are not as effective when separated into individual terms. Pre-translation LCA expansion with single-term concepts can reduce this problem. Pre-translation LCA expansion with single terms is also more effective than pre-translation local feedback and improves both precision and recall. Pre-translation expansion becomes less effective as query disambiguation improves. However, pre-translation expansion terms may still be useful as anchors for disambiguation via the co-occurrence method.

Post-translation LCA is more effective than post-translation local feedback and tends to improve precision. Empirical results show that Combining pre- and post-translation expansion is most effective and improves precision and recall. Tables 6.45 and 6.46 give the recommended configurations based on our empirical results for query expansion via local feedback and LCA, repectively.

Combining either of these two expansion methods with query translation augmented by phrasal translation and co-occurrence disambiguation brings CLIR performance above 90% monolingual without expansion.

Many of the cross-language retrieval comparisons were made with a baseline of monolingual retrieval without expansion. Expansion will obviously raise the baseline. However even with a higher baseline of monolingual with expansion, combining the co-occurrence method with expansion still yields up to 88% of monolingual performance as shown in Table 6.47. This is a considerable improvement over previous work which yielded 68% monolingual.

In the next chapter, we focus on addressing the problem of limited lexical resources. We continue to take a dictionary approach and apply the disambiguation techniques described in this chapter to a *transitive* translation approach. We show

Method	Terms	Documents
Pre-translation (short queries)	5	10-50
Pre-translation (long queries)	10-30	10-20
Post-translation (long queries)	5	10-50
Post-translation (short queries)	20-30	10-50
Combined (short queries)	10-20	30-50
Combined (long queries)	5-20	20-30

Table 6.45. Choice of number of terms and documents or passages for best results when employing local feedback expansion for ambiguity reduction. Combined results achieved by using pre-translation expansion suggestions followed by numbers of terms and documents given in the row labeled "Combined".

Method	Terms	Passages
Pre-translation	20-30	10-20
Post-translation	30-100	30-50
Combined	30	50

Table 6.46. Choice of number of terms and documents or passages for best results when employing expansion via LCA for ambiguity reduction. Combined results achieved by using pre-translation expansion suggestions followed by numbers of terms and documents given in the row labeled "Combined".

Method	Precision	% Monolingual
Mono+exp	0.4113	-
sense1+post	0.3317	81
CO+pre	0.3098	75
CO+combined	0.3533	86
CO+post	0.3623	88

Table 6.47. Average precision as a percentage of that for monolingual with expansion. (Data consists of the TREC-6 AP English corpus and Spanish queries 1,2,4-7, 9-14, 16-24, with Spanish->English translations.)

that transitive translation is a viable means of performing CLIR between languages for which no lexical resource providing a direct mapping exists.

CHAPTER 7

TRANSITIVE TRANSLATION

We have discussed the two main hurdles to effective cross-language retrieval that a CLIR system must address. First, it must find a means of addressing an additional level of ambiguity that arises when trying to map the import of a text object across languages. That is, it must address both within language ambiguity and cross-language ambiguity. Second, it has to incorporate multi-lingual resources that will enable it to perform the mapping across languages. The difficulty here is that there is a limited supply of lexical resources and there are virtually no resources for some pairs of languages.

The goal for addressing ambiguity is to sufficiently reduce its effects such that the gist of the original query is preserved, thus enabling relevant documents to be retrieved. Translation ambiguity is a result of erroneous word translations, failure to translate multi-term concepts as phrases, and the failure to translate out-of-vocabulary words. In Chapter 6, we show how ambiguity greatly reduces the effectiveness of cross-language retrieval in comparison to monolingual retrieval of the same queries. We also show that applying query structure and co-occurrence analysis can greatly reduce the negative effects of translation ambiguity.

This chapter focuses on a dictionary approach to addressing the second hurdle, limited lexical resources. Availability of lexical resources varies and often depends upon several factors including the commercial viability of producing them, proprietary rights, and cost. These problems are described in more detail in Chapter 2.

We continue to take a dictionary approach. More specifically we show that a *transitive* translation approach, where a third language is employed as an interlingua between the source and target languages, is a viable means of performing CLIR between languages for which no bilingual dictionary is available. (Recall that *source* refers to the language being translated and *target* refers to the language being translated into.)

7.1 Experimental Methodology

Our goal here is to develop a means for circumventing the problem of limited availability of linguistic resources. To this end, our approach is applied in the context of transitive translations. In cases where the goal is to perform cross-language retrieval between languages A and C and no bilingual dictionary between A and C is available, transitive translation involves finding a translation path through an intermediate language or *interlingua*. In other words, find a language B for which bilingual dictionaries exist between A and B and between B and C. Language B will act as an interlingua in order to perform translations between A and C.

The main idea behind the following experiments is to simulate a cross-language environment such that the efficacy of a transitive query translation approach can be evaluated and compared to our earlier cross-language work and to monolingual retrieval. In order to do this, we need a set of queries in a source language, a collection of documents in a target language, and relevance judgments. Relevance judgments indicate which documents in the target language are relevant to each query. We can then evaluate the effectiveness of a translation approach by measuring the ability of the translated queries to retrieve relevant documents.

The data for the experiments described herein come from the TREC6 ([129]) cross-language track. There are three sets of 25 queries: one English, one French and one Spanish, where the French and Spanish are manual translations of the English.

Each query set has relevance judgments for a collection of English and a collection of French documents. There are no relevant documents for some of the French queries, so those queries were removed leaving 19 corresponding Spanish and French queries.

In these experiments, two types of translations are performed. One generates French queries from Spanish queries via a bilingual dictionary and we refer to this as a *bilingual translation.* The second type is referred to as a *transitive translation.* Transitive translations are performed from the Spanish queries through English as the interlingua, to French. More specifically, Spanish queries are first translated to English with the Collins Spanish-English machine-readable dictionary. The resulting English query is translated to French via the Collins English-French machine-readable dictionary.

MRD translations are performed as described in Section 4.3. Recall that translation is not used here to suggest the generation of an exact, syntactically correct representation of a query in another language. The words of a query in one language are merely replaced with the dictionary definition of those terms in another language.

Single terms are replaced with only those translation equivalents corresponding to the term's POS. All translation equivalents for an individual term are wrapped in a synonym operator. Words which are not found in the dictionary are added to the new query without translation. Phrasal translations are performed using the dictionary where possible. When a phrase can not be found in the database, it is translated word-by-word and disambiguated via the co-occurrence method discussed in section 6.2.2.5. We do not have a French stemmer, so we employ the Xerox morphological processor to French queries to give the effect of stemming.

Query expansion is applied at various points in the translation process. First it is applied prior to replacement in order to strengthen the query for translation. Queries are also expanded after replacement to reduce the negative effects of erroneous term translations.

7.2 Bilingual VS Transitive Translation Ambiguity

Translation ambiguity greatly reduces cross-language retrieval effectiveness. The following set of experiments is designed to confirm that earlier findings ([12, 13, 14]) about bilingual translations between Spanish and English are also true of bilingual translations between Spanish and French. First, simple bilingual and manual translations to French are generated from the Spanish cross-language queries. Manual translations are generated to simulate near-perfect conditions in which the best translation from the dictionary could always be determined. Later, these results are compared to those for simple transitive translation.

Bilingual translations are generated via word-by-word replacement both with and without the use of POS and synonym operator disambiguation. The two manual translations are performed to measure the degree of ambiguity caused by erroneous word translations and loss of phrasal translations. First is a word-by-word translation where the best single term translation is selected manually. Second is a best word-by-word translation augmented by manual phrasal translation. Table 7.1 compares effectiveness of monolingual retrieval with automatic word-by-word translation with and without POS disambiguation, and with manual word-by-word and manual word-by-word plus phrasal translation.

The bilingual word-by-word translations of the queries achieve 60% monolingual (Table 7.1). This supports earlier reports showing that cross-language retrieval via automatic word-by-word translation without disambiguation achieves 50-60% of monolingual performance. Word ambiguity accounts for 29% of the loss of effectiveness and failure to translate phrases accounts for 40%. This is also consistent with earlier results. However, combining part-of-speech with synonym operator disambiguation is less effective here than reported for previous bilingual translations from Spanish to English. Query term statistics given in Table 7.2 explains at least in part why this is the case. For each query set, the table gives the source and target languages, whether

Query	Mono	WBW	WBW	WBW	WBW+POS	Manual	Manual
			+SYN	+POS	+SYN	WBW	WBW+Phr
Relevant Docs:	1098	1098	1098	1098	1098	1098	1098
Relevant Ret:	730	494	583	533	575	609	605
Avg Prec	0.2767	0.1634	0.2013	0.1869	0.2008	0.1961	0.2419
% Change:		-40.9	-27.3	-32.4	-27.4	-29.1	-12.6
Precision:							
5 docs:	0.5700	0.2900	0.3600	0.3700	0.3600	0.3600	0.4800
10 docs:	0.5050	0.2350	0.2950	0.3000	0.2850	0.2750	0.3900
20 docs:	0.4025	0.1950	0.2425	0.2225	0.2275	0.2300	0.3100
30 docs:	0.3433	0.1717	0.2117	0.1883	0.1983	0.2050	0.2633
100 docs:	0.1850	0.1145	0.1295	0.1230	0.1280	0.1385	0.1510

Table 7.1. Average precision and number of relevant documents retrieved: monolingual French, automatic word-by-word translation, automatic word-by-word translation with SYN, automatic word-by-word translation with POS, automatic word-by-word translation with POS and SYN, manual word-by-word translation, and manual word-by-word with phrasal translation.

POS disambiguation is applied, the average number of words in the translated query, the average number of translation equivalents or definitions per original query term, and the number of original query terms that were recovered via translation. The variances are shown in parentheses.

Source Lang.	Target Lang.	POS	Qry Length	Defs per Term	Undef. Terms	Orig. Terms Recovered
Spanish	English	no	54.67 (738.98)	6.58 (3.5)	4 (6.81)	4.62
Spanish	English	yes	46.52 (447.49)	5.62 (21.43)	4 (6.60)	4.67
Spanish	French	no	16.67 (33.28)	2.11 (3.39)	6 (7.01)	4.19
Spanish	French	yes	15.05 (61.7)	1.92 (2.28)	6 (7.17)	4.14

Table 7.2. Mean (variance) statistics for cross-language query sets: terms per query, definitions per term, undefined terms, number of original query terms recovered after translation.

When bilingual query translations are compared to their monolingual counterparts, each bilingual set contains roughly the same number of original query terms. In other words, the degree to which the monolingual queries are recovered by bilingual translation is roughly the same for translations to French and to English. However, there are more than three times as many translations per query term for bilingual translations between Spanish and English (CSE) than for Spanish to French (CSF) translations. The CSF query terms were replaced with an average of two translations per term while the CSE query terms were replaced with six translations per term. The probability of introducing erroneous terms in translations from Spanish to English is greater. This probability is lower for translations from Spanish to French, and queries contain only 20-30% as many terms. Given the smaller number of French translation equivalents, normalizing for the variance in translation equivalents via the synonym operator has a similar effect to application of POS to reduce the number of translation equivalents. The effects are not additive under these conditions. However, cross-language retrieval effectiveness of CSF queries with automatic translation via a bilingual dictionary and augmented by POS and synonym disambiguation achieves 73% of monolingual effectiveness.

The difference in numbers of translation equivalents for translations from Spanish to English as compared to translations from Spanish to French is likely to be due in part to both the differences in the dictionaries and to the etymology of these languages. First, the Spanish/French dictionary is a pocket version containing approximately 25,000 words, while the Spanish/English dictionary contains roughly 48,000 words. Second, although all three languages belong to the Indo-European branch of languages, English is a member of the Germanic group. French and Spanish belong to the Italic group and are more closely related, therefore share more close cognates. It is interesting to note that although English and French are less closely related, the number of translation equivalents for these queries is similar to that for the more closely related pair, Spanish and French (2.7 and 2.1, respectively). This may be due to the large number of English words borrowed from French, particularly in the domains of government, administration, law, high culture, religion, architecture, the military, and fashion. The TREC6 queries are run on collections of newspaper articles where these topics are common. It would be interesting to see whether we would see different results given different query topics. Additional questions raised regard the effects of working with more closely related versus less closely related languages.

7.3 Transitive Translation Ambiguity

Further analysis reveals that even manual translations introduce ambiguity. There can be many ways to translate a given concept and thus many ways for the original meaning to decay through repeated translations. This suggests that ambiguity will have an even greater negative effect on transitive translations than on bilingual translations.

Table 7.3 shows that this is in fact the case. It compares bilingual translation from Spanish to French to transitive translation from Spanish through English to French. Column one gives recall-precision figures for simple word-by-word bilingual translation via a MRD and employing POS. Column two shows those for a simple word-by-word transitive translation from Spanish through English to French, employing POS. In the transitive case, POS tags from each original Spanish query word are propagated to its English translation equivalents. The English words are then replaced only by the French translation equivalents having the same part-of-speech. Transitive translation effectiveness is 91% below that of the bilingual translation, supporting the assumption that transitive translations are more ambiguous. This makes sense given the statistics in Table 7.2. Each Spanish query word is replaced on average with six English terms. Each English word is replaced on average with 2 French terms. This results in a more ambiguous transitive translation having twelve translation equivalents per original query term.

Query	Bilingual	Transitive
Relevant Docs:	1098	1098
Relevant Ret:	533	216
Avg. Prec.	0.1869	0.0151
% Change:		-91.9
Precision at:		
5 docs:	0.3700	0.0800
10 docs:	0.3000	0.0500
20 docs:	0.2225	0.0400
30 docs:	0.1883	0.0383
100 docs:	0.1230	0.0320

Table 7.3. Average precision and number of relevant documents retrieved for bilingual word-by-word translation and transitive word-by-word translation.

7.4 Resolving Transitive Translation Ambiguity

7.4.1 Synonym Operator

The synonym operator is effective for reducing ambiguity when applied to the target language query following a bilingual translation. We hypothesized that when applied to the inter-lingual translation, it would reduce the additional ambiguity introduced by transitive translation. Table 7.4 gives recall-precision figures that support this hypothesis. Column one shows monolingual retrieval. Column two shows simple word-by-word bilingual translation via a Spanish-French bilingual dictionary. Column three shows simple word-by-word transitive translation from Spanish through English to French. The queries for column four were generated in the following way. First, Spanish query words are replaced by their English translation equivalents after wrapping all equivalents for a particular word in a synonym operator. This creates an English inter-lingual query that ensures that all English translations of a query word are treated equivalently. Then each English word is replaced by its French translation equivalents (recall that the POS for each Spanish query word is propagated to all of its English translation equivalents). The result is that all the French translation equivalents of a particular Spanish query term are treated as an instance of the same word. In other words, it normalizes for the variance in number of French translation equivalents across the original Spanish query words. The effect of ambiguity is considerably reduced, raising effectiveness of transitive translation for fourteen out of 20 queries. This is significant at $p \leq .01$.

Query	Monolingual	Bilingual WBW+POS	Transitive WBW+POS	Transitive WBW+POS +Bi_SYN
Relevant Docs:	1098	1098	1098	1098
Relevant Ret:	730	533	216	328
Avg. Prec.	0.2767	0.1869	0.0151	0.1231
% Change:		-32.5	-94.5	-55.5
Precision at:				
5 docs:	0.5700	0.3700	0.0800	0.2600
10 docs:	0.5050	0.3000	0.0500	0.2300
20 docs:	0.4025	0.2225	0.0400	0.1750
30 docs:	0.3433	0.1883	0.0383	0.1450
100 docs:	0.1850	0.1230	0.0320	0.0775

Table 7.4. Average precision and number of relevant documents retrieved for monolingual retrieval, bilingual word-by-word translation, transitive word-by-word translation, and transitive word-by-word translation with synonym operators used at the bilingual stage to group multiple English translations for a Spanish query term.

Due to the positive effect of applying synonym operators as described above, all subsequent transitive translations are performed in this way.

7.4.2 Phrasal Translation

Failure to translate multi-term concepts as phrases is one of the main factors contributing to translation ambiguity. We hypothesize that this ambiguity will be exacerbated in transitive translations. Consider the following example. The Spanish phrase *Segunda Guerra Mundial* means *Second World War*. The Spanish-English MRD lists (*second, second meaning, veiled meaning, second gear*), (*world-wide, universal, world*), and (*war, warfare, struggle, fight, conflict, billiards*) as the translation equivalents for segunda, mundial, and guerra, respectively. When the English words are translated to French and the French words are grouped such that each group corresponds to one of the original Spanish phrase terms, the phrase translation will be: (*seconde, deuxième, second, licence, avec mention bien assez bien, article de second choix*), (*universel, universelle, du monde, mondial, mondiale*), and (*guerre, guerre,* *lutte, bagarre, combat, lutte, conflit, billard*). The resulting transitive translation of a three word phrase is twenty-seven words long. Furthermore, there are several unrelated terms included in the translation. One means of reducing the effect of this ambiguity is to translate phrases as such when they are listed in the dictionary.

In the next experiment, we translate phrases via the dictionary where possible and where not possible, we disambiguate phrasal translations via the co-occurrence method described in section 6.2.2.5. Spanish queries are first part-of-speech tagged and noun phrases are identified. The query is then translated to English, replacing Spanish phrases with English phrasal translations when they are listed in the dictionary. The resulting English query is translated similarly into French. In addition, we continue to use query structure by applying the passage25, and phrase operators to phrasal translations and by employing the synonym operator for multiple-word query term translations. The synonym operator is also applied to group morphological variants added by the morphological analyzer. For example, *Segunda Guerra Mundial* translates via English phrase dictionary to *Second World War.* This English phrase then translates via French phrase dictionary to *Deuxième Guerre Mondial* and *Second Guerre Mondial.* When the morphological processor is applied, the French translations become *Deuxième Guerre Mondial Mondiale* and *Second Seconder Seconde Guerre Mondial Mondiale*, respectively. This results in each three word phrase being treated as a longer phrase. The synonym operator groups morphological variants to remove this artifact yielding *#passage25(#phrase(Deuxième Guerre #syn(Mondial Mondiale)))* and *#passage25(#phrase(#syn(Second Seconder Seconde)Guerre #syn(Mondial Mondiale))).*

Table 7.5 compares monolingual retrieval with bilingual translation and transitive translation both with phrasal translation via dictionary and co-occurrence disambiguation. Transitive translation effectiveness increases by 15.7% when phrases are translated this way. In this particular query set, nine of twenty-one Spanish phrases were translatable to English via dictionary. Of those nine English phrases, five were translatable to French via dictionary. The remaining phrases were translated via co-occurrence disambiguation.

Query	Mono	Bilingual WBW+Phr+Co	Transitive WBW+Phr+Co
Relevant Docs:	1098	1098	1098
Relevant Ret:	730	596	389
Avg. Prec.	0.2767	0.2104	0.1424
% Change:		-23.9	-48.6
Precision at:			
5 docs:	0.5700	0.3500	0.2700
10 docs:	0.5050	0.3000	0.2300
20 docs:	0.4025	0.2425	0.1800
30 docs:	0.3433	0.2067	0.1567
100 docs:	0.1850	0.1325	0.0860

Table 7.5. Average precision and number of relevant documents retrieved for monolingual, bilingual word-by-word translation with phrase dictionary and co-occurrence translation of phrases, and transitive word-by-word translation with phrase dictionary and co-occurrence translation of phrases.

When translations are disambiguated only via synonym operators and phrasal translation, transitive translation achieves 51% monolingual while bilingual translation with the same disambiguation strategies achieves 76% monolingual. Transitive translation is 32% less effective than bilingual translation. Recall from Table 7.2 that the average number of English definitions for a Spanish query term is about six and that there are two French definitions per English query term. This means that transitive translations have about 12 translations per query term. Although there are far more translations per query term in the transitive translation, the number of original query terms in the resulting queries is about the same. Transitive translation yields more ambiguous queries than bilingual translation.

Earlier work [14] showed that the effectiveness of bilingual translations could be brought near the level of monolingual. The next section explores the feasibility of reducing transitive translation ambiguity by first generating the best bilingual translation. This means reducing as much inter-lingual ambiguity as possible prior to the transitive translation phase.

7.4.3 Combining Disambiguation Strategies at the Bilingual Phase of a Transitive Translation

In the following experiments, all disambiguation strategies discussed in section 6.2 and [14] are applied at the bilingual stage of the transitive translations. In other words, when translating from Spanish, all disambiguation strategies are applied to generate the least ambiguous English query possible. This disambiguated English query is then used as the base for a transitive translation to French.

Bilingual translations are generated via automatic dictionary translation augmented by co-occurrence disambiguation and query expansion. Expansion is applied both prior to and after translation to English. While these approaches work well to reduce the negative effects of ambiguity on retrieval for one level of translation, this effect may not be preserved after an additional level of translation and further introduction of ambiguity. The mean number of interlingua terms after the bilingual translation to English with the application of all disambiguation techniques is 126.8, while the original query has 7.19. The mean number of original query terms recovered via translation in the bilingual queries is 5.86. Although many of these additional terms aid in disambiguation, there are more than one hundred more query terms after bilingual translation than are in the original query. It is not clear what effect this will have on the transitive translation from English to French.

Dictionaries often list a phrase or short description as the translation of a head word, especially if the head word is a verb. For example, the translation for *rodear* meaning “to surround” includes: *to beat about the bush, to go by an indirect route,* and *to make a detour.* Many of these types of translations will not be listed as head words in a dictionary. Translating these multi-term definitions word-by-word may introduce even more ambiguity. To test this possibility, two transitive translations were performed beginning with the bilingual translations as a base; one in which these multi-term definitions were translated word-by-word and one in which they were translated via co-occurrence disambiguation. Table 7.6 gives recall-precision values for the results from these two transitive translation approaches and compares them to transitive translation without applying expansion at the bilingual translation phase and to monolingual retrieval. This is to establish whether the expansion at the bilingual stage adds noise or improves the final translation. (i.e. Translate Spanish queries to English via MRD with synonym operator, phrase dictionary and co-occurrence phrasal translation, both with and without any expansion. Translate the resulting queries from English to French.)

Query	Monolingual	Transitive No Biling. Expan.	Transitive +Biling. Expan. +WBW	Transitive +Biling. Expan. +Co
Relevant Docs:	730	389	669	679
Avg. Prec	0.2767	0.1424	0.1845	0.1784
% Change:		-48.5	-33.3	-35.5
Precision at:				
5 docs:	0.5700	0.2700	0.3000	0.3000
10 docs:	0.5050	0.2300	0.2650	0.2700
20 docs:	0.4025	0.1800	0.2375	0.2350
30 docs:	0.3433	0.1567	0.2183	0.2217
100 docs:	0.1850	0.0860	0.1490	0.1465

Table 7.6. Average precision and number of relevant documents retrieved for monolingual retrieval, transitive translation with no bilingual expansion, transitive translation after bilingual expansion and word-by-word translation of multi-term definitions, and transitive translation after bilingual expansion and co-occurrence disambiguation of multi-term definitions.

As one would expect, there is a significant improvement (at $p \leq .019$) in effectiveness when we generate the best inter-lingual translation possible prior to performing the transitive translation. However, there is no significant difference between effectiveness of queries translated with and without co-occurrence disambiguation of multi-term inter-lingual definitions. This could be due to the following explanation.

We have an ambiguous interlingua translation that contains erroneous translations as well as synonymous correct translations. When we attempt to further disambiguate, we reduce the expansion effect generated by correct synonymous translations. Expansion has been shown to reduce the effect of ambiguity. In addition, there may be some natural disambiguation that occurs when many related and unrelated terms come together. One would expect that more related or correct than unrelated or incorrect terms would co-occur in documents and thus reduce the effect of poor translations.

Results also show that the reduction in ambiguity as exhibited by the bilingual translations to English is not lost after an additional level of translation. Transitive translation effectiveness is 67% of monolingual French. This is reasonable when one considers that the bilingual translations of the Spanish queries having only one level of translation achieved 76% monolingual. Bilingual translations to English gained further improvements after query expansion. The following experiments explore whether transitive translations can be further disambiguated via query expansion.

7.4.4 Transitive Translations and Query Expansion

Transitive translations are more ambiguous than bilingual translations. Post-translation expansion has been shown to significantly reduce the effects of bilingual translation ambiguity. We attempted to apply it (results not shown) at the final stage of transitive translations to ascertain whether it is effective for further reduction of transitive translation ambiguity. Contrary to our expectations, results did not show any further reduction of ambiguity when expansion was applied in this way. Query-by-query analysis reveals why this is so.

The expansion terms selected by LCA are unrelated to the query content. Query one is about the controversy over Waldheim's WWII activities, but expansion terms are about agriculture, wine growing, and other unrelated concepts. In fact, only query 13 appears to have any related expansion terms. Query 13 asks about the attitudes of the Arabic countries towards the peace process in the Middle East. The expansion terms include *caire, syrie,* and *arafat.* It is unusual, judging by previous experience as well as work published by others using LCA, for expansion concepts to be unrelated to the query topic. LCA typically selects better expansion terms than do other expansion techniques.

In addition to seeing no increase in effectiveness when transitive translations were expanded, our initial French monolingual expansions showed no increase in effectiveness. Further investigation suggests that the lack of effectiveness of French monolingual expansion was due in part to a programming error in the LCA software that was used. More specifically, the software did not deal appropriately with empty documents existing in the French database. The inclusion of empty documents as input to LCA negatively affected the way in which expansion terms were selected. This resulted in many unrelated expansion terms being added to the query. By the time this error was discovered, other aspects of the environment had changed sufficiently that the earlier experiments could not be redone.

After removal of the empty documents, we re-ran French monolingual expansion experiments. Table 7.7 gives recall precision figures for LCA expansion of French monolingual queries after the removal of the empty documents from the LCA analysis.

The best expansion runs were generated by the addition of 100 and 200 terms from the top 50 passages. A Wilcoxon signed-ranks test shows these results are statistically significant ($p \leq 0.11$ and $p \leq 0.04$, respectively) and show improvement in retrieval effectiveness of approximately 5% over unexpanded monolingual queries. They show clearly that expansion can be effective for French queries.

Our French monolingual expanded queries showed a smaller increase in effectiveness than Spanish, English, and Chinese monolingual expanded queries have shown in the past [133]. Monolingual LCA expansion of Spanish, English, and Chinese showed effectiveness increases of 13%, 30%, and 14%, respectively. The question that remains

Query	Monolingual	LCA50-100	LCA50-200
Relevant:	1098	1098	1098
Rel_ret:	730	759	756
Avg prec:	0.2767	0.2901	0.2891
% Change:		4.8	4.5
Precision:			
5 docs:	0.5700	0.5800	0.5900
10 docs:	0.5050	0.5150	0.5300
20 docs:	0.4025	0.4250	0.4150
30 docs:	0.3433	0.3583	0.3550
100 docs:	0.1850	0.2005	0.1950
R-Precision (precision after R (= num_rel for a query) docs retrieved):			
Exact:	0.3167	0.3311	0.3227

Table 7.7. Recall-precision pairs, R-precision, and number of relevant documents retrieved for expansion of monolingual French queries. These results were generated after removing empty French documents from the LCA analysis. The best results were achieved by selecting the top 100 or 200 LCA expansion terms from the top 50 passages.

is whether this difference is related to the nature of French, or whether it is system related. It seems unlikely that it would be the former. Query expansion has been applied successfully, in a monolingual environment, to a number of languages including Spanish, Chinese, and Japanese ([3, 4, 58]). Although the French cross-language work by Buckley [23] also failed to improve after expansion, they did apply expansion to their monolingual French runs. However, there is no report of how much an improvement was realized over monolingual without expansion. In addition, in work by Boughanem and Soulé-Dupuy [17], expansion increased monolingual French retrieval effectiveness by 11%, providing more evidence that the nature of French is not an obstacle to the application of expansion.

One difference between our monolingual French runs and those mentioned above is that they employed a simple stemmer. Stemming has been shown to improve retrieval effectiveness, which could impact the effectiveness of expansion. Rather than stem, we apply the XEROX morphological processor, which has been shown to work as well as a traditional stemmer for English [66]. Our monolingual French runs were 43% more effective with morphological processing than without. However we do not have stemmed monolingual French runs for comparison and this still does not answer the question of whether expansion is more effective when stemming is employed.

Although it is unlikely that there are characteristics of the language that would prevent improvements in retrieval effectiveness after expansion of French queries, effectiveness may be dependent upon LCA parameter settings. As discussed in Section 1.3, it is not unusual for strategies for applying IR techniques to vary across languages. Xu's [133] work showed that the parameters at which the best improvements were gained for LCA expansion of English, Spanish, and Chinese differed with respect to parameter settings. If our French monolingual runs were repeated with optimal parameter settings, we may see further improvements in effectiveness.

7.5 Summary and Discussion

In Chapter 6 we showed that statistical techniques are effective for reducing the ambiguity associated with bilingual query translation. The experiments described herein support those results. More importantly, we have demonstrated that transitive retrieval via machine readable dictionaries is possible. Transitive retrieval effectiveness is 67% monolingual, which compares favorably with bilingual translation at 76% monolingual. Although there is still room for improvement, we have learned several positive lessons about the feasibility of transitive translation. First, although transitive translation is much more ambiguous than bilingual translation, that ambiguity can be significantly reduced. Table 7.8 illustrates this. It compares bilingual translation without expansion to transitive translation without expansion at the bilingual stage and transitive translation after application of all ambiguity reduction techniques at the bilingual stage. In other words, Spanish queries are translated to English using the synonym operator, POS and co-occurrence disambiguation, and combined expansion. The resulting English query is then translated to French. This brings transitive translation effectiveness up 30% to 67% of monolingual effectiveness. Although the results are not directly comparable, this is still as good or better than effectiveness reported for these queries via other cross-language approaches based on more complex resources that mapped the source and target languages directly.

Query	Bilingual Without Expansion	Transitive No bilingual Expansion	Transitive With bilingual Expansion
Relevant Docs:	1098	1098	1098
Relevant Ret:	596	389	669
Avg Prec	0.2104	0.1424	0.1845
% Change:		-32.4	-12.3
Precision:			
5 docs:	0.3500	0.2700	0.3000
10 docs:	0.3000	0.2300	0.2650
20 docs:	0.2425	0.1800	0.2375
30 docs:	0.2067	0.1567	0.2183
100 docs:	0.1325	0.0860	0.1490

Table 7.8. Average precision and number of relevant documents retrieved for bilingual translation without expansion, transitive translation without expansion at the bilingual stage and transitive translation with expansion at the bilingual stage.

Second, transitive translations can be as effective or more effective than their monolingual or bilingual counterparts. Nine queries each of the transitive and bilingual translations are more effective than monolingual, while ten and eleven queries, respectively, are less effective than monolingual. Eleven of the transitive translations are more effective than their bilingual counterparts.

Table 7.9 gives query term statistics for the monolingual, bilingual, and transitive query sets (these statistics were collected after morphological processing). What it shows is that although the transitive translations are longer and more ambiguous, they still recover more unique original query terms (54.8%) than do bilingual translations (45%). This suggests that it may be possible to combine evidence from transitive translations via several intermediate languages to further reduce ambiguity associated with this approach.

Our results suggest that transitive translation is a viable approach, but there are still open questions. We were only able to re-run the French monolingual expansion

Type of Translation	Source Language	Target Language	Qry Length	Original Terms	Terms Recovered
Monolingual	French	French	13.76 (28.56)	N/A	N/A
Bilingual	Spanish	French	17.0 (49.9)	7.1 (16.47)	45%
Transitive	Spanish	French	459.2 (21145.6)	8.33 (23.4)	54.8%

Table 7.9. Mean (variance) query term statistics for French monolingual, bilingual, and transitive translations: terms per query, number of original query terms recovered by translation, and percentage of unique query terms recovered by translation.

experiments. We suspect that expansion of French bilingual and transitive translations would yield further improvements, but it has yet to be shown. It also remains to be seen whether changing parameter settings for the French monolingual expansions would yield further improvements.

In addition to issues related to differences in strategies for applying expansion and other techniques to different languages, there are still many issues to explore regarding the impact that lexical resources and languages have in this environment. More work must be done, for example, to try to determine how the relatedness of source, target, and interlingua languages or how the quality of the machine readable dictionaries influence effectiveness.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

In recent years there has been tremendous growth in the amount of multi-lingual, electronic media. The current boom in economic growth has lead more businesses and organizations to expand the boundaries of their organizational structure to include foreign offices and interests. Greater numbers of people are interested in data and information collected in other regions of the world as efforts to address issues of global concern lead to increased multi-national collaboration. In addition, the explosive growth of the Internet and use of the World Wide Web (WWW) makes possible access to information that was once hindered by physical boundaries. The Internet has become a powerful resource of information. This growth in availability of multi-lingual data in all areas of the public and private sector is driving an increasing need for systems that facilitate access to multi-lingual resources by persons with varying degrees of expertise with foreign languages. Cross-language retrieval technology is a means of addressing this need.

Cross-language retrieval aims to develop tools that in response to a query posed in one language (e.g. Spanish), allow the retrieval of documents written in other languages (e.g. Chinese). Unlike a machine translation system, the goal of a CLIR system is not to generate exact, syntactically correct representations of a text in other languages. It is rather to cull through the tremendous number of electronic texts and to select and rank those documents that are most likely related to a query written in another language.

This dissertation describes the problems encountered in a cross-language retrieval environment. A dictionary-based approach to CLIR is presented and shown to be effective; as effective in fact as approaches relying on more complex techniques and resources. This chapter reviews the primary contributions of the work and then discusses directions for future investigation.

8.1 Research Contributions

There are two main obstacles that make cross-language retrieval a difficult problem. The first is that whenever one deals with language, one is confronted with ambiguities caused by the imprecise ways in which ideas are expressed. This is relatively easy for a human to decipher, but quite difficult to do automatically. The second is the scarcity of lexical resources. Most resources are a product of manual effort whose generation is dependent upon the needs of a well-funded organization willing to pay for its production or upon commercial incentive. This dissertation investigates both of these problems and contributes to the advancement of information retrieval in four main ways.

- Analysis of the causes of ambiguity associated with cross-language retrieval.

- Development of practical and effective statistical techniques for CLIR which rely as little as possible on scarce resources.

- Comparative analysis of different approaches to the task showing, contrary to popular belief, that effectiveness is not dependent upon complex linguistic analysis.This suggests that CLIR systems can be adapted to many languages with little effort, unlike MT systems which require a significant effort for each new language pair.

- A general, effective approach to CLIR via machine readable bilingual dictionaries which provides the foundation of a general architecture for effective CLIR in a general domain. Most importantly, the transitive approach suggests a means for circumventing the problem of limited resources.

The following subsections describes each of these contributions in more detail.

8.1.1 Ambiguity Analysis

In order to perform cross-language retrieval, information in one language must be transferred to another language. Both within-language ambiguity and between-language ambiguity must be addressed so that the gist of a text object is preserved. Cross-language retrieval effectiveness is generally 40-60% below that of monolingual retrieval. We have shown that translation ambiguity is a result of erroneous word translations, failure to accurately translate multi-term concepts, and out-of-vocabulary words.

Words are ambiguous because they can have multiple parts-of-speech, each of which may have multiple meanings. In a cross-language environment this is exacerbated by the fact that even if we resolve within language ambiguity by identifying the intended meaning of a word, it can be non-trivial to map it to its correct representation in another language. Our analyses show that word ambiguity accounts for up to 50% of cross-language retrieval error (the difference between monolingual and cross-language retrieval effectiveness).

OOV terms are those words for which no dictionary translation is available. Important query words that can not be expressed in the target language will obviously have a negative impact on effectiveness. We found that OOV words can account for more than 20% of retrieval error.

Multi-word constructs such as phrases and collocations are responsible for up to 31% of CLIR error. The difficulty here is two-fold. First, phrases can not always be translated word-by-word. Second, the quality of the phrasal translation is very important. The negative impact of a poor phrasal translation can be great enough to further degrade CLIR effectiveness. Our work suggests it may be better to replace some phrases word-by-word, than to choose a poor phrasal translation.

8.1.2 General and Effective Ambiguity Reduction Techniques

Despite the negative impact of translation ambiguity on cross-language retrieval effectiveness, we develop techniques that can be applied to greatly reduce the effects of ambiguity. These techniques are based on query structure, syntactic analysis, or statistical analysis of word co-occurrence. Prior to this work being done, it was commonly believed that resources more complex than a bilingual dictionary would be necessary to enable effective CLIR. We now know that by combining the ambiguity reduction strategies described below, the effectiveness of CLIR via MRD in a general domain can be brought near the level of monolingual retrieval. This is considerably better than the 40-60% monolingual effectiveness achieved by automatic dictionary translation alone.

8.1.2.1 Synonym Operators and Part-of-speech

Erroneous words are the main source of retrieval error in a CLIR environment. This occurs for two reasons. First, because dictionaries often include archaic usages for head-words, query translations can be dominated by these rarely used equivalents. Second, dictionary translations contain many related words (some of which may not be appropriate to the query context). Query words having the greatest number of translation equivalents tend to be given greater importance. This is an artifact of document relevance assessment since the words that a query and document have in common serve as the basis for inferring the likelihood that they are related.

Query structure has been employed in other areas of information retrieval, primarily to make more explicit the way in which words are being used or to emphasize how important a word or group of words is (are) to the concept being conveyed. This is the first time it has been applied to directly address translation ambiguity. The synonym operator is simple to apply and significantly reduces the effects of ambiguity.

Without the synonym operator, query effectiveness can be reduced by two factors. First, infrequent translations (e.g. archaic usages) get more weight than more frequent translations due to their higher idf. Second, a query word with n times as many translation equivalents is treated as n times more important. The synonym operator treats occurrences of all words within it as occurrences of a single pseudo-term thus having a disambiguating effect by de-emphasizing infrequent words. In addition, it reduces ambiguity by normalizing for the variance in number of translation equivalents across query terms.

When queries are expressed in well-formed sentences, part-of-speech tagging can be applied. A term to be translated can then be replaced with only those translation equivalents corresponding to its correct POS. Part-of-speech tagging has been used by others to guide selection of translation equivalents, however until now assessment of its effectiveness in a CLIR environment has not been done. We find that although increases in average precision for a query set in which POS is combined with other ambiguity strategies suggest it is effective, when used alone it did not produce significant improvements for any of our query sets. This is good news since people rarely express a query as a syntactically correct sentence. This is especially important in a web environment where average query length is two words.

8.1.2.2 Co-occurrence Analysis of Unaligned Corpora

A word typically takes its meaning from the context in which it is used. Examining the ways that words occur together tells us something about what the author of a text object intended to express. The application of co-occurrence analysis to lexical disambiguation is not new. However, it has widely been assumed to be useful, so exploration of its effectiveness in a specific environment has been infrequent. In information retrieval, co-occurrence analysis has primarily been applied to the vocabulary mismatch problem via query expansion or to measure association between text objects. In a cross-language environment it has been employed primarily with aligned corpora for translation term selection. Techniques based on parallel and comparable corpora exploit the known associations between documents in those corpora to infer probable translations. A common belief is that this knowledge gives some advantage over approaches employing unaligned corpora. We show this to be untrue and show that aligned corpora are not necessary. More specifically, we show that co-occurrence analysis can be successfully employed with unaligned corpora for both the selection of terms in phrase translation and to reduce the effects of translation ambiguity.

Query expansion reduces the effects of translation ambiguity in two ways. First, when applied prior to query translation, it strengthens the query by the addition of content specific terms which act as translation anchor points. This tends to boost precision, and to a lesser extent, recall. Second, when applied after translation, expansion diminishes the negative impact of poor translations and improves recall. When combined, improvements mediated by pre- and post-translation expansion are additive and significantly reduce the effects of ambiguity.

Co-occurrence analysis of unaligned corpora provides a means for exploring patterns of word usage from which to infer the likelihood that a set of words is used in a specific context. We apply this technique to sets of words comprising alternative translation equivalents for compositional phrases and show that the *em* measure is a viable means of selecting good phrasal translations. In addition, preliminary work employing unambiguous anchor terms suggests it to be a viable means of selecting translation equivalents for single words. We also show via comparative analysis that co-occurrence disambiguation using unaligned corpora is as good as, if not better than disambiguation via a parallel corpus.

8.1.3 Effectiveness Does Not Require MT-type Linguistic Analysis

When this line of research was first started, many believed that the way to effective cross-language retrieval was via machine translation. This is partially to do with an alternate view of the nature and goals of information retrieval in general and of the nature and goals of CLIR in particular. An information retrieval system must estimate the degree to which documents in a collection reflect the information expressed in a user query. This is done by generating representations of queries and documents that can then be compared. One might assume that this requires natural language understanding and complex natural language processing tools. However, the impact of NLP techniques on IR has not been great enough to outweigh the cost of employing them. Most effective IR systems do not include processing beyond simple POS tagging and phrase formulation. Because the language gap must be automatically addressed by a CLIR system, one tends to think of machine translation. A machine translation system's goal is given a text, to generate an exact, syntactically correct representation in another language. However, the goal of a CLIR system is to cull through large amounts of textual data and to select those documents that are most likely related to the query. Like a monolingual system, a CLIR system will reach its goal by generating and comparing representations of documents and queries. Despite the language gap, it is not necessary that these representations be strict translations. It is sufficient that they be approximate "translations" that convey the gist of the original text object. This is good news, given that current MT systems provide low- to medium-quality translations for arbitrary input and given the complexity of building an MT system for a new language pair.

We compared our dictionary-based approach to CLIR with two commercial machine translation systems, SYSTRAN and Langenscheidt's T1. More specifically, we performed retrieval on the queries translated by the MT systems and compared retrieval results with those of our dictionary-based approach. We found that cross-language retrieval via our approach is as effective or more so than the application of machine translation. In fact, in our study the dictionary approach was 18% more effective than machine translation via SYSTRAN.

8.1.4 Foundation for General Architecture for CLIR

We have identified the problems confronting the task of cross-language retrieval and presented a number of techniques for addressing them. We applied these techniques to a cross-language approach based on automatic dictionary translation and showed that effective CLIR can be achieved without linguistic analysis or machine translation. Our translations were performed between Spanish, English and French. However, the disambiguation techniques are general enough to be applied to other languages as well as to approaches relying on resources other than dictionaries. In addition, the research is carried out using the INQUERY information retrieval system, but the techniques presented in this thesis are general enough to be used in other IR systems. This dictionary-based approach is effective and provides the foundation for a general architecture for cross-language information retrieval as given by Figure 8.1.

Building a CLIR system requires tools for addressing language specific characteristics such as character representation, language identification, word morphology, and language typography. Character encodings vary across and within languages. English is easily represented via the ASCII code, which is a standard readily processed by most systems. However other languages based on the Roman alphabet, or on non-Latin alphabets, or based on a combination of several alphabets have additional orthographic characters typically represented by a byte code larger than 128. These may require special treatment for input, indexing, and display. There may be several ways to encode these character sets and all of them must be supported. The language in which the query is written could be specified by the system-user, or alternatively, identified automatically. Language identification has been applied to both on-line

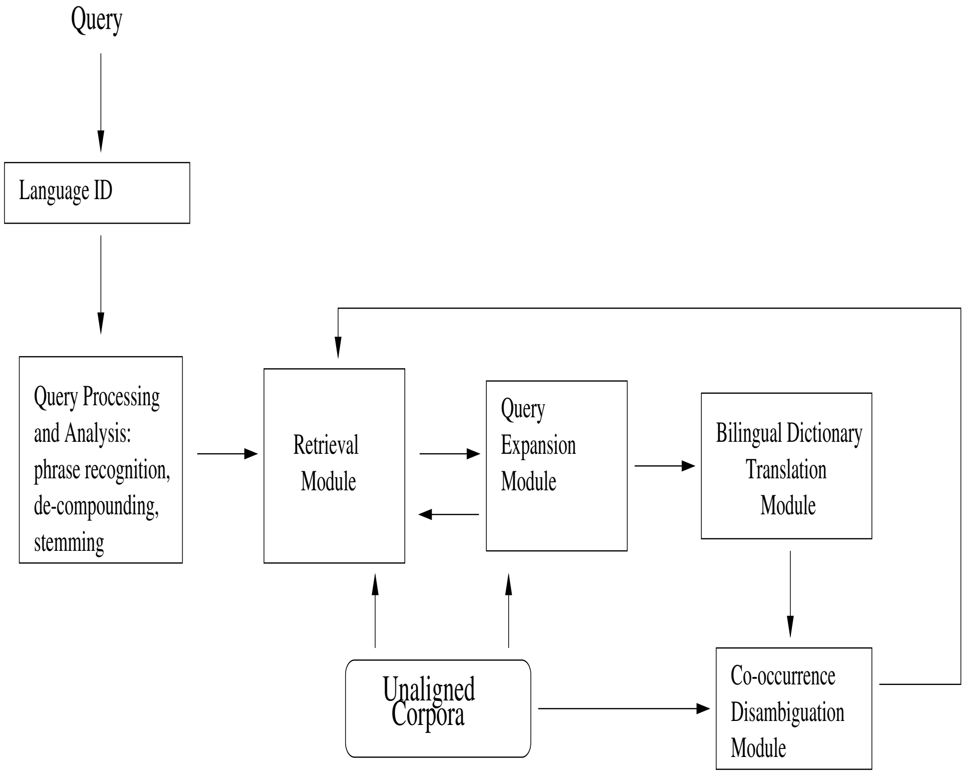
Query
Language ID
Query Processing and Analysis: phrase recognition, de-compounding, stemming
Retrieval Module
Query Expansion Module
Bilingual Dictionary Translation Module
Unaligned Corpora
Co-occurrence Disambiguation Module

Figure 8.1. General architecture for cross-language retrieval based on Machine Readable Dictionaries.

text [43] and to images converted to text [115], with accuracy ranging from 92% with as little as ten characters of input to > 99% with 50 or more characters of input. In addition, the system may incorporate modules to perform stemming, part-of-speech tagging, word segmentation, and compound splitting.

Translation of phrases is crucial to cross-language retrieval effectiveness. Therefore, a module for phrase recognition must be implemented. Language-dependent recognizers combining part-of-speech tagging and syntax rules are possible, but work by Mitra et. al [91] suggests language specificity is not necessary. In Mitra's work, statistical phrase recognition is performed via identifying any pair, triple, quadruple, etc of non-stop words that appear adjacent to one another. If such a grouping occurs in at least 25 documents, the group is selected as a phrase. The study showed that syntactic and statistical methods of phrase recognition yielded comparable performance on high-precision retrieval.

The degree to which the system must rely upon language-specific tools will vary. There is disagreement among researchers regarding the effectiveness of certain techniques such as stemming, so some mono- and multilingual systems do not employ it. However, other techniques may be more important for some languages. Take Dutch, for example, which is rich in compounding. Kraaij [75] reports that applying a word-splitter to compounds improves retrieval performance. Many language specific tools such as stemmers and part-of-speech taggers are now readily available. Moreover, research continues to be done in the area of multilingual retrieval where a system is tailored for the language upon which it operates. Rather than re-inventing the wheel, cross-language retrieval systems should incorporate language specific tools developed and applied in multilingual environments.

The foundation of our cross-language approach is the application of language-independent techniques for ambiguity reduction. More specifically, modules to support structured query operators including the synonym operator, and query expansion techniques are crucial for emphasizing query context and for reducing the impact of poor translations. Additionally, modules supporting co-occurrence analysis of unaligned corpora should be employed. Co-occurrence analysis reduces ambiguity further by facilitating the selection of translation equivalents for query constituents such as compositional phrases. The corpora that are searched for relevant documents provide the context for translation equivalent selection. In order to support co-occurrence techniques applied prior to translation, the system should also include unaligned corpora in all languages for which it will accept queries.

Bilingual and/or multilingual dictionaries provide the mapping between lexical items across languages. CLIR based on machine-readable dictionaries is significant for two reasons. First, there are MRDs for many commercially-important languages. Resources such as aligned corpora are less prevalent. Second, MRDs have fewer of the disadvantages associated with other multilingual resources. MRDs are more readily available, less expensive, and require less work to prepare than other resources. Furthermore, dictionaries provide broader coverage than other resources making them more applicable in a general domain. This implies that a CLIR system based on MRDs could be quickly re-targeted to some new language pairs merely by the addition of a new dictionary.

Results indicate that CLIR via transitive MRD translation is a viable approach. This increases the significance of MRD-based cross-language retrieval by making retrieval possible between languages for which no direct translation resources are available. In other words, we can perform retrieval between two languages even if there are no available resources that explicitly provide word correspondences.

8.2 Future Work

This chapter discusses several areas for further investigation that would logically follow the work in this dissertation. These areas can be divided into two types. The first is work that explores more deeply, questions related to experiments presented in this dissertation. The second is work that explores areas that were not addressed herein. Sections 8.2.1, 8.2.2, 8.2.3, 8.2.4, 8.2.5 and 8.2.6 discuss the former and the remaining sections discuss the later.

8.2.1 Impact of the Degree of Language Relatedness

We have shown that a cross-language approach based on bilingual and transitive translations are viable. The work in this dissertation is based upon translations between three languages belonging to the Indo-European branch of languages. Spanish and French are more closely related to each other than either of them are to English. However, all three of these languages are more closely related than any one of them is to Arabic or Chinese, for example. We still need to explore the effect of language relatedness on this approach and on how that may effect strategies for applying ambiguity reduction techniques.

8.2.2 Impact of Dictionary Quality and Coverage on Effectiveness

Although dictionaries may be more prevalent and simpler to apply than resources such as parallel or comparable corpora, the available resources exhibit a wide range in quality. On one end of the spectrum are on-line word lists created cooperatively by visitors to host sites, who voluntarily add entries. This approach provides a large pool of would-be lexicographers, however quality control is difficult to implement. On the other end of the spectrum are commercially produced dictionaries developed by well-trained staff. The quality of dictionary entries is good in these environments and several versions of the same resource are often developed. These versions typically differ in breadth of coverage since the number of headwords varies between versions.

The work in this dissertation employed commercially produced dictionaries. Our dictionaries differed in the average number of translation equivalents per headword. Word-by-word translations between Spanish and French performed better relative to monolingual retrieval than did WBW translations between Spanish and English. This appears to be due to less ambiguous Spanish/French translations since they contain roughly one-third as many equivalents as their Spanish/English counterparts. This may be due to a difference in the mapping of conceptual frameworks between languages or it may be an artifact caused by a difference in the depth of headword entries between the two dictionaries. In either case, a better understanding of the relationship between these factors and ambiguity is desirable.

Moreover, our dictionaries have about 25,000, 30,000, and 60,000 entries for the Spanish/French, English/French, and Spanish/English versions, respectively. The next question that remains is whether a system using lower quality dictionaries and/or word lists can achieve cross-language effectiveness at the level of our experiments. If performance is degraded by lower quality resources, then it would be useful to be able to quantify the impact of resource quality on effectiveness. This would enable system builders to focus more effort and capital on those languages for which high quality results are most critical.

8.2.3 Evaluation of Statistical Phrase Recognition

In our experiments, phrase identification was performed by first part-of-speech tagging terms and then grouping them via syntactic rules. This approach is applicable when the input text is grammatically correct, as it is in our document collections. However, queries are typically not expressed in grammatically correct sentences. Since phrasal translation can be critical to query effectiveness, another method of phrasal identification is needed. Statistical phrase recognition has been shown to be comparable to syntactically-based methods and one would expect this to be the case here. Nevertheless, it must be shown empirically to be so.

8.2.4 Guiding the Application of Ambiguity Reduction Techniques

Each of the techniques described in this work has been applied to all queries and average performance was evaluated. More often than not, a query by query analysis shows the techniques improve effectiveness. However, not every query is improved and the factors impacting effectiveness remain unclear. For example, ambiguity reduction via both the synonym operator and query expansion may be related to both source query length and the number of translation equivalents per query term. It is well-known that in a monolingual environment, IR-techniques are sensitive to retrieval variables. Further exploration of the impact of query properties on effectiveness of ambiguity reduction strategies would help to identify their limitations. In addition, this knowledge may provide heuristics for guiding their application on a query-by-query basis.

8.2.5 Effectiveness with Non-Western Languages

The choice of languages and test collections was driven by the availability of TREC resources. Our results suggest that our dictionary-based approach is general enough to be applied across Latin-based Western languages. Many of the techniques in our approach have been successfully employed in other environments with non-Western languages such as Japanese. Although we suspect this to also be the case when applied for the purpose of cross-language ambiguity reduction, showing this for a language such as Chinese would strengthen the viability of our approach.

8.2.6 Translating OOV Terms

In our work, "pseudo-translations" of queries were generated via machine readable dictionary and then augmented via statistical ambiguity reduction strategies. We did not attempt to directly address the out-of-vocabulary word problem. When a query word could not be found in the dictionary, it was added to the new query without being translated. OOV terms can be a significant problem to cross-language effectiveness. This will be especially damaging when an OOV term is the main aspect of a query, so strategies for effectively translating them must be developed. One area of investigation is the applicability of transitive translation to this problem. It is possible that multiple transitive translations may provide enough evidence for identification and selection of OOV term translation equivalents. Moreover, analysis of the translation equivalents of words co-occurring with OOV terms in unlinked corpora may provide additional evidence for their successful translation.

8.2.7 Data Fusion

Our cross-language approach is effective because implicit relationships between words and concepts in different languages were identified at the query level. A related question is how one might identify such relationships across documents as is necessary for data fusion; more specifically, relationships between documents written in different languages.

The data fusion task is to infer the most likely ranking when combining retrieval results from different query representations or document collections. Different query representations or text collections retrieve different sets of relevant and non-relevant documents. Comparing the retrieved sets is difficult, particularly if distributed collections use different internal representations. Rules for combining these results take advantage of document similarities such as word usage.

Cross-language documents are retrieved using query representations in different languages (e.g. French and English) from different monolingual databases. In addition to problems related to distributed, heterogeneous databases, cross-language documents have fewer explicit similarities. We believe that the documents retrieved across languages are related and therefore that the words across documents are also related. However, it's possible that few or no documents retrieved in one language are relevant. The problem is finding ways to increase our belief in the relatedness of objects across foreign language documents to aid in combining them appropriately. Moreover, techniques for comparing documents in different languages may influence our understanding of the ways in which other distinct data types are compared. For example, it is not obvious how to compare different multi-media objects.

8.2.8 Vocabulary Switching

The vocabularies of different domains often use different words to describe the same concept, theme, or idea. For this reason, when a person searches for information outside of her regular domain of discourse, there may be little overlap between the query words used and the vocabulary of potentially relevant documents. The computational complexity of many of the previous approaches to this problem place limits on the contexts in which they can be applied.

Techniques for identifying explicit relationships in multi-lingual information may suggest a more general approach for automatically recognizing connections between sub-languages. More specifically, further analysis of the relationships identified by co-occurrence techniques in the cross-language environment may be applicable to identifying object relationships across sub-languages. Co-occurrence techniques that uncover semantic relationships could be applied to a variety of information technologies including automatic thesaurus generation, browsing, and categorization.

BIBLIOGRAPHY

[1] Activa dictionaries and translation service. http://www.activa.arrakis.es (July 1998).

[2] Alis technologies and the internet society: Web languages hit parade. http://babel.alis.com:8080/palmares.html (June 1997).

[3] Allan, J., Ballesteros, L., Callan, J., Croft, W.B., and Lu, Z. Recent experiments with inquery. In *Proceedings of the Fourth Retrieval Conference (TREC-4) Gaithersburg, MD: National Institute of Standards and Technology* (1995).

[4] Allan, J., Callan, J., Croft, W.B., Ballesteros, L., Broglio, J., Xu, J., and Shu, H. Inquery at trec-5. In *Proceedings of the Fifth Retrieval Conference (TREC-5) Gaithersburg, MD: National Institute of Standards and Technology* (1996).

[5] Allan, James, Callan, Jamie, Croft, W. Bruce, Ballesteros, Lisa, Byrd, Don, Swan, Russell, and Xu, Jinxi. Inquery does battle with trec-6. In *Proceedings of the Sixth Retrieval Conference (TREC-6) Gaithersburg, MD: National Institute of Standards and Technology* (1997).

[6] AltaVista-SYSTRAN. Babelfish. http://www.infotektur.com/demos/babelfish/en.html, July 2000.

[7] Arnold, Doug, Balkan, Lorna, Meijer, Siety, Humphreys, R.Lee, and Sadler, Louisa. *MACHINE TRANSLATION: An Introductory Guide.* Blackwells-NCC,London, 1994, pp. 201–205.

[8] Attar, R., Choueka, Y., Dershowitz, N., and Fraenkel, A.S. Kedma - linguistic tools for retrieval systems. *Journal of the Association for Computing Machinery 25*, 1 (January 1978), 52–66.

[9] Attar, R., and Fraenkel, A. S. Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery 24* (1977), 397–417.

[10] ALLwords Multi-Lingual Search Dictionary. http://www.allwords.com/ (1998).

[11] Baker, Mark. Complex predicates and agreement in polysynthetic languages. In *To appear: Complex Predicates*, A. Alsina, J. Bresnan, and P. Sells, Eds., CLSI, Stanford, pp. 249–290.

[12] Ballesteros, Lisa, and Croft, W. Bruce. Dictionary-based methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications* (1996), pp. 791–801.

[13] Ballesteros, Lisa, and Croft, W. Bruce. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval* (1997), pp. 84–91.

[14] Ballesteros, Lisa, and Croft, W. Bruce. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st International Conference on Research and Development in Information Retrieval* (1998), pp. 64–71.

[15] BBN part-of-speech tagger for spanish. http://www.gte.com/bbnt/ (July 1999).

[16] A web of on-line dictionaries.
http://www.facstaff.bucknell.edu/rbeard/diction1.html#multi (Jan. 1998).

[17] Boughanem, M., and Soulé-Dupuy, C. Mercure at trec6. In *Proceedings of the Sixth Retrieval Conference (TREC-6) Gaithersburg, MD: National Institute of Standards and Technology* (1997), pp. 321–328.

[18] Braschler, M., and Schauble, P. Multilingual information retrieval based on document alignment techniques. In *Second European Conference on Research and Advanced Technology for Digital Libraries* (1998), pp. 183–197.

[19] Brill, E. A simple rule-based part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing* (Trento, Italy, 1992).

[20] Broglio, J., Callan, J., and Croft, W.B. Inquery system overview. In *Proceedings of the TIPSTER Text Program (Phase I)* (1994), pp. 47–67.

[21] Brown, P., Cocke, J., Pietra, S. Della, Pietra, V.J. Della, Jelinek, F., Lafferty, J.D., Mercer, R.L., and Roossin, P.S. A statistical approach to machine translation. *Computational Linguistics* (1990).

[22] Brown, Peter F., DellaPietra, Stephen A., DellaPietra, Vincent J., and Mercer, Robert L. Word sense disambiguation using statistical methods. In *Proceedings 29th Annual Meeting of the Association for Computational Linguistics* (Berkeley, CA, 1991), pp. 265–270.

[23] Buckley, Chris, Mitra, Mandar, Walz, Janet, and Cardie, Claire. Using clustering and superconcepts within smart: Trec6. In *Proceedings of the Sixth Retrieval Conference (TREC-6) Gaithersburg, MD: National Institute of Standards and Technology* (1997), pp. 107–121.

[24] Callan, J., and Croft, W. B. An evaluation of query processing strategies using the tipster collection. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1993), pp. 347–356.

[25] Callan, J.P., Croft, W.B., and Broglio, J. Trec and tipster experiments with inquery. *Information Processing and Management 31*, 3 (1995), 327–343.

[26] Carbonell, J. G., Yang, Y., Frederking, R. E., Brown, R., Geng, Y., and Lee, D. Translingual information retrieval: a comparative evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'97)* (1997).

[27] Church, Kenneth. Char_align: A program for aligning parallel texts on character level. In *Proceedings of the Annual Meeting of the ACL* (1993), pp. 1–8.

[28] Bilingual machine readable dictionaries. Electronic Reference, HarperCollins Publishing, Ltd.

[29] Croft, Bruce, Callan, J., and Broglio, J. Trec-2 routing and ad-hoc retrieval evaluation using the inquery system. In *Proceedings of the Second Text Retrieval Conference (TREC-2) Gaithersburg* (1994), Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-215.

[30] Croft, Bruce, Turtle, Howard R., and Lewis, David D. The use of phrases and structured queries in information retrieval. In *Proceedings of the Fourteenth Annual ACM/SIGIR Conference on Research and Development in Information Retrieval.* (1991), A. Bookstein, Y. Chiaramella, G. Salton, and V. V. Raghavan, Eds., ACM Press.

[31] Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. A practical part-of-speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy.* (1992), pp. 133–140.

[32] Dagan, Ido, and Church, Kenneth W. Termright: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Natural Language Processing (ANLP-94)* (1994), pp. 34–40.

[33] Dagan, Ido, Itai, Alon, and Schwall, Ulrike. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (1991), pp. 130–137.

[34] Daniel, Wayne W. *Applied nonparametric statistics.* Boston: Houghton Mifflin, 1978.

[35] Davis, Mark. New experiments in cross-language text retrieval at nmsu's computing research lab. In *Proceedings of the Fifth Retrieval Conference (TREC-5) Gaithersburg, MD: National Institute of Standards and Technology* (1996).

[36] Davis, Mark, and Dunning, Ted. Query translation using evolutionary programming for multi-lingual information retrieval. In *Proceedings of the Fourth Annual Conference on Evolutionary Programming* (1995).

[37] Davis, Mark, and Dunning, Ted. A trec evaluation of query translation methods for multi-lingual text retrieval. In *Proceedings of the Fourth Retrieval Conference (TREC-4) Gaithersburg, MD: National Institute of Standards and Technology, Special Publication 500-236* (1995).

[38] Davis, Mark W., and Ogden, William C. Quilt: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval* (1997), pp. 92–98.

[39] Dutch dictionary project.
http://www.freedict.com/dictionary/index.html (January 1998).

[40] De Mente, Boye Lafayette. Asian business codewords: Killing with silence. http://www.apmforum.com/columns/boye36.htm (March 2000).

[41] The dict development group. http://www.dict.org/ (February 1999).

[42] Dixie multilingual on-line dictionary. http://www.cs.ut.ee/g̃ordon/dixie/dixie.cgi (January 1998).

[43] Dunning, Ted. Statistical identification of language. Tech. rep., CRL, New Mexico State University, 1994. Technical Memor MCCS-94-273.

[44] Dunning, Ted, and Davis, Mark. Multi-lingual information retrieval. Technical report MCCS-93-252, Computing Research Laboratory, New Mexico State University, 1993.

[45] Eurodicautom. http://eurodic.echo.lu/cgi-bin/edicbin/EuroDicWWW.pl (January 1998).

[46] Fujii, H., and Croft, W.B. A comparison of indexing techniques for japanese text retrieval. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1993), pp. 237–247.

[47] Fujii, H., and Croft, W.B. Comparing the retrieval performance of english and japanese text databases. In *2nd Annual Workshop on Very Large Corpora* (1994).

[48] Fung, P., and McKeown, K. A technical word and term translation aid using noisy parallel corpora across language groups. *Machine Translation* (1996), 53–87.

[49] Furnas, G.W., Deerwester, S., Dumais, S.T., anmd R.A. Harshman, T.K. Landauer, Streeter, L.A., and Lochbaum, K.E. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th International Conference on Research and Development in Information Retrieval* (1988), pp. 465–480.

[50] Gachot, Denis A., Lange, Elke, and Yang, Jin. An application of machine translation technology in multilingual information retrieval. In *Cross-Language Information Retrieval*, Gregory Grefenstette, Ed. Kluwer Academic Publishers, 1996, ch. 9, pp. 105–118.

[51] Gale, W.A., and Church, K.W. Identifying word correspondences in parallel text. In *Fourth DARPA Workshop on Speech and Natural Languages* (1991), pp. 152–157.

[52] Gale, William, and Church, Kenneth. A program for aligning sentences in bilingual corpora. *Computational Linguistics 19* (1993), 75–102.

[53] Gerber, Laurie, and Yang, Jin. Systran MT dictionary development. Tech. rep., SYSTRAN Software, Inc., 1997.

[54] Global Internet Statistics (by language), Euro-Marketing. http://www.euromktg.com/globstats (May 1998).

[55] Gms, gesellschaft fuer multilinguale systeme. http://www.gmsmuc.de/english/ (Jan. 1998).

[56] Graff, David, and Finch, Rebecca. Multilingual text resources at the linguistic data consortium. In *Proceedings of the 1994 ARPA Human Language Technology Workshop* (1994).

[57] Guthrie, Joe A., Guthrie, Louise, Wilks, Yorick, and Aidinejad, Homa. Subject-dependent co-occurrence and word sense disambiguation. In *Proceedings of the 29th Conference of the Association for Computational Linguistics* (1991), pp. 146–152.

[58] Han, C., Fujii, H., and Croft, W.B. Automatic query expansion of japanese text retrieval. UMASS Technical Report, 1994.

[59] Hansard Corpus. Linguistic Data Consortium. http://morph.ldc.upenn.edu/Catalog/LDC95T20.html (July 2000).

[60] Harman, D. How effective is suffixing? *Journal of the American Society for Information Science 42*, 1 (1991), 7–15.

[61] Harman, D. Overview of the third text retrieval conference (trec-3). In *Proceedings of the Third Text REtrieval Conference* (1995), D. Harman, Ed., pp. 1–20.

[62] Harman, D. Overview of the fourth text retrieval conference (trec-4). In *Proceedings of the Fourth Text REtrieval Conference* (1996), D. Harman, Ed., pp. 1–23.

[63] Harman, Donna, Ed. Proceedings of the 4th Text Retrieval Conference. 1995.

[64] Hearst, M. A. Improving full-text precision on short queries using simple constraints. In *Symposium on Document Analysis and Information Retrieval (SDAIR)*, University of Nevada, Las Vegas.

[65] Hollander, Myles, and Wolfe, Douglas A. *Nonparametric statistical methods*, 2nd ed. New York: J. Wiley, 1999.

[66] Hull, D. Stemming algorithms - a case study for detailed evalutation. *Journal of the American Society for Information Science 47* (1996), 70–84.

[67] Hull, David. A weighted boolean model for cross-language text retrieval. In *Cross-language Information Retrieval*, Gregory Grefenstette, Ed. Kluwer Academic Publishers, 1998, pp. 119–136.

[68] Hull, David A., and Grefenstette, Gregory. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* (1996), pp. 49–57.

[69] Hutchins, W. John, and Somers, Harold L. *An Introduction to Machine Translation.* Academic Press Limited, 1992.

[70] The internet dictionary project. http://www.june29.com/IDP/ (Jan. 1998).

[71] International Conference on New Methods in Language Processing. *Probabilistic Part-of-Speech Tagging Using Decision Trees* (1994).

[72] Jeff, Moad. Machine translation–the next generation. In *PC Week*. January 1998.

[73] Kittredge, Richard I. *Machine translation. Theoretical and methodological issues.* Cambridge University Press, 1987, ch. 4, p. 64.

[74] Klavans, Judith L., and Tzoukermann, Evelyne. Combining corpus and machine-readable dictionary data for building bilingual lexicons. *Machine Translation 10* (1996).

[75] Kraaij, Wessel. Viewing stemming as recall enhancement. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* (1996), pp. 40–48.

[76] Kraaij, Wessel, and Hiemstra, Djoerd. Cross language retrieval with the twenty-one system. In *Proceedings of the Sixth Retrieval Conference (TREC-6) Gaithersburg, MD: National Institute of Standards and Technology* (1997), pp. 753–760.

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

[78] Krovetz, R., and Croft, W. Bruce. Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems 10*, 2 (1992), 115–141.

[79] Kupiec, Julian M. An algorithm for finding noun phrase correspondances in bilingual corpora. In *Proceedings, 31st Annual Meeting of the ACL* (1993), pp. 17–22.

[80] Kwok, K.L. Comparing representations in chinese information retrieval. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval* (1997), pp. 34–39.

[81] Landauer, Thomas K., and Littman, Michael L. Fully automatic cross-language document retrieval. In *Proceedings of the Sixth Conference of the UW Center for the New Oxford English Dictionary and Text Research* (1990), pp. 31–38.

[82] *Larousse De Poche Dictionnaire Français Espagnol/Español Francés.* Larousse Bordas, 1997.

[83] *Longman Dictionary of Contemporary English.* Longman, 1978.

[84] Lee, Joon Ho, and Ahn, Jeong Soo. Using n-grams for korean text retrieval. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* (1996), pp. 216–224.

[85] Lernout & Hauspie. L&H *i*Translator. http://www.lhsl.com/itranslator/demo/, July 2000.

[86] Lesk, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, 1986.

[87] Loether, Herman J., and McTavish, Donald G. *Descriptive and inferential statistics: an introduction*, 2nd ed. Boston: Allyn and Bacon, 1980.

[88] Matsumoto, Y., Kurokashi, S., and Myoki, Y. *User's Guide for the JUMAN System - A User-Extensible Morphological Analyzer for Japanese.* Nagao Lab, Kyoto University, 1991.

[89] McCarley, J. Scott, and Roukos, Salim. Fast document translation for cross-language information retrieval. In *Machine Translation and the Information Soup. Proceedings of the Third Conference of the AMTA* (1998), David Farwell, Laurie Gerber, and Eduard Hovy, Eds., Springer, pp. 150–157.

[90] Miller, George. Wordnet: An on-line lexical database. *Journal of Lexicography 3*, 4 (1990).

[91] Mitra, M., Buckley, C., Singhal, A., and Cardie, C. An analysis of statistical and syntactic phrases. In *Conference Proceedings of RIAO-97* (1997), pp. 200–214.

[92] Nagao, Makoto. *Machine Translation. How Far Can it Go?* Oxford University Press, 1989.

[93] Nirenburg, Sergei. *Machine Translation: Theoretical and methodological issues.* Cambridge University Press, 1987, ch. 1.

[94] Oard, Doug, and Diekema, Anne. Cross-language information retrieval. In *Annual Review of Information Science And Technology (ARIST)*, Martha Williams, Ed., vol. 33. Information Today Inc., Medford, NJ, 1998.

[95] Oard, Douglas, and Hackett, Paul. Document translation for cross-language text retrieval at the university of maryland. In *Proceedings of the Sixth Retrieval Conference (TREC-6) Gaithersburg, MD: National Institute of Standards and Technology* (1997), pp. 687–696.

[96] An index of on-line dictionaries. http://www.artinternet.fr/city/Biblio/Autres/autres.htm (June 1998).

[97] Paxson, Vern. Flex: Fast lexical analyzer generator. Tech. rep., Lawrence Berkely Laboratory, California, 1995.

[98] Peter Schäuble, Páraic Sheridan. Cross-language information retrieval (clir) track overview. In *Proceedings of the Sixth Text REtrieval Conference (TREC-4)* (1998), E.M. Voorhees and D.K. Harman, Eds., NIST Special Publication 500-240, pp. 31–41.

[99] Peters, Carol, and Picchi, Eugenio. Across languages, across cultures: Issues in multilinguality and digital libraries. *D-lib magazine, The Magazine of Digital Library Research* (May 1997).

[100] Picchi, Eugenio, and Peters, Carol. Cross language information retrieval: A system for comparable corpus querying. In *Cross-Language Information Retrieval*, Gregory Grefenstette, Ed. Kluwer Academic Publishers, 1996, ch. 7, pp. 81–92.

[101] Pirkola, Ari. The effects of query structure and dictionary setups in dictionary-base cross-language information retrieval. In *Proceedings of the 21st International Conference on Research and Development in Information Retrieval* (1998), pp. 55–63.

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

[103] Radwan, K. Vers l'Accés Multilingue en Langage Naturel aux Bases de Données Textuelles. *PhD thesis, Université de Paris-Sud, Centre d'Orsay.* 1994.

[104] Resnik, Philip. Selectional preference and sense disambiguation. In *Proceedings of the ANLP-97 Workshop: Tagging Text with Lexical Semantics: Why, What, and How?* (Washington, D.C., 1997).

[105] Rijsbergen, C. J. Van. *Information Retrieval.* Boston: Butterworth Publishers Inc., 1975.

[106] Robertson, S. E. The probability ranking principle in ir. *Journal of Documentation 33* (77), 294–304.

[107] Salton, G., and McGill, M.J. *Introduction to Modern Information Retrieval.* Butterworth Publishers Inc., 1983.

[108] Salton, Gerard. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Publishing Company Inc., 1983.

[109] Salton, Gerard, and Buckley, Chris. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science 41* (1990), 288–297.

[110] Sanderson, Mark. Word sense disambiguation and information retrieval. In *Proceedings of the 17th ACM SIGIR Conference* (1994), pp. 142–151.

[111] Sanderson, Mark. *Word Sense Disambiguation and Information Retrieval.* PhD thesis, Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK, 1997.

[112] Schütze, Hinrich, and Pedersen, Jan. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval* (1995), pp. 161–175.

[113] Sheridan, Paraic, and Ballerini, Jean Paul. Experiments in multilingual information retrieval using the spider system. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* (1996), pp. 58–65.

[114] Sheridan, Paraic, Braschler, Martin, and Schauble, Peter. Cross-language information retrieval in a multilingual legal domain. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries* (1997), pp. 253–268.

[115] Sibun, P., and Reynar, J. Language identification: Examining the issues. In *Symposium on Document Analysis and Information Retrieval* (1996), pp. 125–135.

[116] Singal, A., Buckley, C., and Mitra, M. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1996), pp. 21–29.

[117] Smadja, Frank, McKeown, Kathleen R., and Hatzivassiloglou, Vasileios. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics 22*, 1 (1996), 1–38.

[118] Stillman, Robert E. *The new philosophy and universal languages in seventeenth-century England: Bacon, Hobbes, and Wilkins.* Lewisburg: Bucknell University Press; London:Cranbury, NJ: Associated University Presses, 1995.

[119] SYSTRAN Software, Inc. http://www.systranet.com (Jan. 1998).

[120] SYSTRAN Software, Inc. SYSTRAN architecture. www.systrasoft.com/mt_arch.htm (Dec. 1999).

[121] Tucker, Allen B. *Machine translation. Theoretical and methodological issues.* Cambridge University Press, 1987, ch. 2, pp. 30–31.

[122] Turtle, Howard R., and Croft, W. Bruce. Efficient probabilistic inference for text retrieval. In *RIAO 3 Conference Proceedings* (1991), pp. 664–661.

[123] Turtle, Howard R., and Croft, W. Bruce. Inference networks for document retrieval. In *Proceedings of the 14th International Conference on Research and Development in Information Retrieval* (1991), pp. 1–24.

[124] UN Corpus. Linguistic Data Consortium. http://morph.ldc.upenn.edu/Catalog/LDC94T4A.html, July 2000.

[125] van der Eijk, Pim. Automating the acquisition of bilingual terminology. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics* (1993), pp. 113–119.

[126] van Rijsbergen, C. J. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation 33* (1977), 106–119.

[127] Voorhees, Ellen M. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (1993), pp. 171–180.

[128] Voorhees, E.M., and Harman, D.K., Eds. Proceedings of the 5th Text Retrieval Conference (TREC-5). 1996.

[129] Voorhees, E.M., and Harman, D.K., Eds. Proceedings of the 6th Text Retrieval Conference (TREC-6). 1997.

[130] Wein, Charlotte. Nine problems concerning arabic. Tech. Rep. 166-CAT(WS)-9-E, Center for Contemporary Middle East Studies, Odense University, Denmark, 1995.

[131] Wu, Z., and Tseng, G. Chinese text segmentation for text retrieval achievements and problems. *JASIS* (October 1993).

[132] Xerox finite-state morphological analyzers. http://www.xrce.xerox.com:80/research/mltt/Tools/morph.html (Dec. 1998).

[133] Xu, Jinxi. *Solving The Word Mismatch Problem Through Automatic Text Analysis.* PhD thesis, Department of Computer Science at the University of Massachusetts, Amherst, Amherst, Massachusetts. USA, 1997.

[134] Xu, Jinxi, and Croft, W. Bruce. Querying expansion using local and global document analysis. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval* (1996), pp. 4–11.

[135] Xu, Jinxi, and Croft, W. Bruce. Corpus-based stemming using co-occurrence of word variants. *To appear in ACM TOIS January* (1998). Technical Report TR96-67, Dept. of Computer Science, University of Massachusetts/Amherst.

[136] Yang, Jin, and Gerber, Laurie. Chinese-english machine translation system. In *Proceedings of the International Conference on Chinese Computing* (1996).

[137] Yarowsky, David. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (1995), pp. 189–196.

[138] ZERES Dictionary Online. http://www.zeres.de/dict/ (January 1998).