A Machine Translation Approach to Cross Language Text Retrieval

by

María Gabriela Fernández-Díaz

ISBN: 1-58112- 267-5

DISSERTATION.COM

Boca Raton, Florida
USA • 2005

To my family…

Acknowledgments

I would like to thank first my supervisor Mark Sanderson for his guidance and support throughout my project and writing of my thesis. I am also grateful to many members of the Computing Science Department at the University of Glasgow, especially Keith van Rijsbergen, Mark Dunlop and Richard Cooper, who have contributed to my understanding of modern information systems.

Finally I would like to thank my friends and family for supporting me during this year.

Abstract

Cross Language Text Retrieval (CLTR) has been defined as the retrieval of documents in a language different from that of the original query. To make this possible some kind of mechanism has to be applied in order to translate the information contained in the source sentence. Many different approaches have been carried out with the purpose of transferring the information from the source language query to the target language one. Though all these methods deal with a way of translating as much information as possible from the source query, little research has been conducted in relation to the field of Machine Translation (MT). The purpose of this research work is to determine the feasibility of using MT techniques for CLTR. Specifically, I will describe how a MT system has been adapted without much effort to translate Spanish queries of a specific domain, i.e. Finance and Economics, into English in order to retrieve documents related to that field. The results of this process will then be compared with the results obtained from the retrieval of the original English queries. Thus, I will discuss the advantages and disadvantages of using MT for CLTR.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Cross language text retrieval means the retrieval of documents based on queries, regardless of the language in which the documents and the query are expressed. Because the monolingual text retrieval problem has been well studied, the emphasis here is put on the cross-language aspect of text retrieval. That is, the case in which the queries are expressed in a language different from that of the documents. Researchers, when faced with the dilemma of translating the whole sets of document collections or the sets of queries, usually choose the second option, as queries by nature are shorter than documents. For this reason, the literature in this field is filled with samples of all the different methods employed so far for translating queries.

Among these methods, the dictionary-based, which make use of machine-readable dictionaries, have proved the most popular, as the task of simply transferring all the senses of an entry from the source query to the target one does not imply a great effort. However, results gained via this approach suggest that the information contained in such dictionaries is not enough, causing significant drops in the retrieval effectiveness of these systems when compared to the performance of their monolingual counterparts. To solve this problem, new approaches should be explored. One of these involves the use of machine translation systems. Machine translation involves the sophisticated use of natural language processing and text generation techniques. It is characterized by the direct resolution of ambiguity in translation using structural information from the source

language text. This strategy for cross language text retrieval might allow the researcher to take advantage of the extensive body of research on machine translation and the availability of commercial products. However, it is generally argued that the performance of current machine translation systems in the setting of general language translation is not good enough yet to make this option entirely satisfactory. One weakness of present fully automatic machine translation systems is that they are able to produce high quality translations only in limited domains. Though some research has already been conducted in this direction, not many decisive results have been presented. This phenomenon alone encouraged me to explore the way machine translation systems could be better approached for their use in a cross language text retrieval environment. This dissertation constitutes thus a first approach to the matter. Some hints and conclusions are presented for future broader research in this area.

1.2 Thesis Aim

The main aim of this thesis is to determine the feasibility of using machine translation techniques for cross-language text retrieval. In order to decide about the role machine translation plays in the retrieval of text across languages, different experiments will be performed in this thesis. The experimentation in this research involves the use of a machine translator as well as an information retrieval system. In the first case, a limited domain machine translation system, Lekta, will be adapted for its use in a cross-language text retrieval context. This system was originally developed for its use in a banking environment. Thus, the subject field in Lekta is restricted to those sentences commonly used for banking transactions. In the second case, SIRE, an information retrieval system will be used so that the evaluation and results of the experiments can be drawn.

To measure the feasibility of using machine translation techniques, two goals will be pursued in this work. On the one hand, the linguistic costs of adapting Lekta's parser,

the analyser of the system, will be measured. Thus, we could decide whether the benefits we get in the retrieval process are worthy of these costs or not. On the other hand, the results we will obtain when submitting machine translated queries will be compared with the results obtained from the original queries. Therefore, the effectiveness of machine translation systems when retrieving across languages will be examined. Having these goals in mind, different kinds of experiments will be performed throughout the thesis. Because this work represents a first approach to the subject matter being discussed, the number of experiments will help us to decide which are the advantages and disadvantages of using machine translation for this particular research. In this sense, we will be able to draw some conclusions and decide about the role of machine translation for future work.

1.3 Thesis Outline

Chapter two introduces the field of Information Retrieval (IR). First of all, some of the concepts related to the field that appear in this work are briefly explained. Then, I will focus my attention on the field of Cross Language Text Retrieval (CLTR). At this point, I will describe in detail the three main approaches commonly related to CLTR. That is to say, knowledge, corpus and text translation-based approaches. Experiments and results carried out following these methods will be presented.

Chapter three is devoted to the field of Machine Translation (MT). The way MT systems work will be explained, so that a better understanding of the field can be achieved. Furthermore, the different types of current systems are detailed. In the second part of the chapter, I will describe Lekta; a MT tool developed to translate automatically sentences in a banking context from English into Spanish and vice versa. In addittion, Lekta's architecture will be explained both from a linguistic and a computational point of view.

Chapter four lays the foundations for the experimentation. First of all, the goals pursued in this dissertation are detailed. Secondly, a description of the queries and documents selected for the experiments is presented. Thirdly, we will show the way Lekta has been adapted for the translation of the new queries. In this sense, we will explain the adaptation of the lexicon of the parser, on the one hand, and some of the changes introduced in the grammar of the analyser, on the other. Finally, some examples will be offered, including the translations of some of the commonest queries.

Chapter five constitutes the final part of the project, that is, the evaluation. In this chapter, translations obtained from the machine translation system will be submitted to SIRE, the information retrieval system. Thus, we will be able to evaluate the results obtained. On the other hand, we will make a comparison between the results obtained when submitting original queries and the ones obtained when submitting translated ones and we will draw some conclusions.

Chapter six concludes summarizing the work presented in this thesis. Furthermore, it will highlight some of the main points of this research and will point out the final conclusions.

Chapter seven, finally, presents some of the limitations of the work carried out in this project and gives some hints for future work in this area.

Chapter 2

Cross Language Information Retrieval

2.1 Introduction

The main purpose of this chapter is to give a general overview of the current state of the art in Cross Language Text Retrieval (CLTR). First of all, I will give a general introduction to the field of Information Retrieval (IR). Secondly, I will focus my attention on a subfield of IR, that is, CLTR. Thus, I introduce some definitions that have been recently proposed to refer to this discipline. Afterwards, I will establish the main basis for the motivation of this research topic. Then, some of the main trends in CLTR are presented. In this sense, I will talk about the three main approaches commonly related to CLTR, i.e., knowledge, corpus and text translation-based. Finally, I will conclude this chapter discussing the main points of this research work and presenting my future research in this dissertation.

2.2 Information Retrieval

Text is the primary way that human knowledge is stored. Techniques for storing and searching for textual documents have been used for years. Computers, however, have changed the way text is stored, searched, and retrieved. With automated systems, the number of indexing terms that can be used for an item becomes much larger. The

subfield of computer science that deals with the automated storage and retrieval of documents is called information retrieval (IR). IR shares concerns with many other computer subdisciplines, such as artificial intelligence, multimedia systems, parallel computing, and human factors.

In an IR system, the user enters a query, either formulated in a structured language or as a natural language phrase, and the system replies with all documents contained in a document database that match the query. This section introduces some of the basic concepts in IR that will be presented in this work. Thus, three main subsections are presented. The first one refers to the process of indexing, one of the main steps that takes place in an IR system. The second one talks about the weighting of terms and the third one about the typical way the evaluation of the results is presented in most IR systems. By no means, this section tries to give an exhaustive review of the state of the art in IR. For a deeper study of the field, see Salton and McGill (1983), Van Rijsbergen (1979) and Frakes (1992).

2.2.1 Indexing

Generally speaking, the purpose of indexing is to produce a representation of the queries and documents in a form suitable for a computer to use. Although in the early years of IR, the task of indexing was performed manually by human indexers, modern IR employs numerous content analysis techniques to automate this process. These techniques fall under two broad categories: they can be either statistical or linguistic. While statistical text analysis uses word frequency information, the linguistic approach to text analysis uses natural language processing or other artificial intelligence techniques to exploit the syntactic and semantic relationships of the words in a given context. Although over the years statistical analysis has proven its usefulness, there is still a controversy on whether linguistic methods can be of real use to IR. The rest of this section gives a brief description of the indexing process, as seen from the statistical viewpoint.

The first step towards obtaining a set of document representatives (or descriptors) is to remove from the input text all high-frequency words (usually called *stop-words*), as their contribution to the discriminating power of the descriptors is minimal, and their removal can reduce the text volume up to 50 percent[1]. Categories such as articles, adverbs and prepositions form part of the set of stop-words.

The next step consists of removing word suffixes and prefixes so that each word is reduced to its stem. This process, that improves retrieval effectiveness and reduces the size of the indexing files, has been referred to as *stemming*. Stemming will cause the words 'engineer', 'engineered' and 'engineering' to be reduced to the common stem 'engine'. However, stemming will also cause the words 'retrieve' 'retrieval', and 'retriever' to be reduced to the common stem 'retriev', disregarding the fact that the first two words should be distinguished from the third. Nonetheless, it is taken for granted that the number of such errors does not have a real effect on the retrieval effectiveness. Furthermore, it is generally believed that stemming either improves the retrieval effectiveness or has no effect at all. For an overview of some of the systems that make use of stemming algorithms as well as discuss different models of implementation, including the well-known Porter stemming algorithm (Porter, 1980), see Frakes (1992).

Each of the stems detected from a document can represent one of its descriptors. Descriptors are also known as *index terms*. Once word stems have been generated as possible index terms, the weighting process follows in order to identify those stems with an important role for content identification.

2.2.2 Weighting

The weighting process assigns a *weight* to each index term, according to its importance for content identification. Most weighting methods are based on the observation that the frequency of occurrence of a word in a text is related to its importance for content

[1] See Van Rijsbergen (1979), pp. 18-19.

representation. What characterizes a word as a useful index term, is the fulfilment of the following requirements (Salton and McGill, 1983):

- It must be related to the information content of the document so as to render the item retrievable when it is wished, thus having a great proportion of the relevant documents retrieved. The proportion of the relevant documents that are retrieved is known as *recall*.

- It must distinguish the documents to which it is assigned from the remainder to prevent the indiscriminate retrieval of all items, whether wanted or not, thus having a great proportion of the retrieved documents to be relevant. The proportion of the retrieved documents that are relevant is known as *precision*.

Terms with high frequency of occurrence in a document seem to be useful for the first requirement. This suggests the usage of a *term frequency* (*tf*) factor as the first part of the weighting scheme.

Terms with low frequency of occurrence in the whole collection seem to be useful for the second requirement. This suggests the usage of a within document frequency factor, also known as *inverse document frequency* (*idf*), as the second part of the weighting scheme.

Using the product of the term frequency tf_{ij} and the inverse document frequency idf_j for a term j of a document i, one can obtain a measure of the importance of that term for content identification of that document, by using the following weight w_{ij} for that term (Salton and Yang, 1973):

$$W_{ij} = tf_{ij} \cdot idf_j$$

The simplest form of the *tf* component is the binary one: *tf* is equal to 1 for a term present in a document and 0 is equal to the others. A set of other possible forms of the *tf* component can be found in [Salton and Buckley, 1988]. A typical form follows:

$$tf_{ij} = FREQ_{ij}$$

where $FREQ_{ij}$ is the raw term frequency (number of times the term *j* occurs in the document *i*).

A typical *idf* component may be computed as (Sparck Jones, 1972):

$$idf_j = \log (N / n_j)$$

where n_j is the total number of occurrences of term *j* in the collection and *N* the number of the documents in the collection.

The $tf \cdot idf$ scheme is frequently normalized when the length of the document collections is not homogeneous. The normalized $tf \cdot idf$ weighting is defined in Salton and Buckley (1988).

2.2.3 Evaluation

IR systems can be evaluated in terms of many criteria including execution efficiency, storage efficiency, retrieval effectiveness, and the features they offer to a user. Execution efficiency is measured by the time it takes a system, or part of a system, to perform a computation. Storage efficiency is measured by the number of bytes needed to store data. Most IR experimentation has focused on retrieval effectiveness, usually based on document *relevance judgements*. Thus, documents retrieved in response to a query are judged by the users as being relevant or not. This has been a problem since relevance

judgements are subjective and unreliable. That is, different judges will assign different relevance values to a document retrieved in response to a given query. A detailed discussion of the issues involved in IR experimentation can be found in Salton and McGill (1983) and Sparck-Jones (1981).

Many measures of retrieval effectiveness have been proposed. The most commonly used are *recall* and *precision,* which have been introduced implicitly during the description of weighting methods in the previous section. Recall (*R*) is the proportion of relevant documents retrieved for a given query over the number of relevant documents for that query in the database. Except for small test collections, this denominator is generally unknown and must be estimated by sampling or some other method. Precision (*P*) is the proportion of the number of relevant documents retrieved over the total number of documents retrieved.

For the rank position *i* of each relevant document in response to a query, precision and recall values are calculated:

$$P_i = r_i/n_i,\ R_i = r_i/Rel$$

Where Rel is the number of relevant documents, r_i is the number of relevant documents returned at that point of rank position *i* and n_i is the number of documents returned at that point of rank position *i*. The *P* and *R* values for each query are interpolated and then the average values are calculated for all the queries in order to have a set of precision values at recall points of 0.1, 0.2, … 1, from which graphs are constructed.

For a detailed analysis of various IR evaluation methods one can refer to Salton and McGill (1983) and Van Rijsbergen (1979).

2. 3 Defining Cross-Language Text Retrieval

This section gives a general overview of the state of the art in cross language text retrieval (CLTR), a discipline that is growing quickly as more and more information in different languages can be easily accessed from networked systems. Before going further, I will explain the meaning of the idea implied by CLTR.

CLTR means the retrieval of documents based on queries using natural language, regardless of the language in which the documents and the query are expressed. Because the monolingual text retrieval problem has been well studied, the emphasis here is put on the cross-language aspect of text retrieval. That is, the case in which queries are expressed in a language different from that of the documents.

At the SIGIR' 96 workshop on "Cross-Linguistic Information Retrieval" the participants discussed the proliferation of terminology being used to describe the field and settled on "Cross-Language" as the best single description of the salient aspect of the problem (Oard, 1997a). "Multilingual" was felt to be too broad, since that term has also been used to describe systems able to perform within-language retrieval in more than one language but that lack any cross-language capability. The terms "cross-lingual" and "cross-linguistic" were also felt to be equally good descriptions of the field, but "cross-language" was selected as the preferred term in the interest of standardisation and it is for this reason that I have chosen this term for doing my research work.

Regarding the term "text retrieval", I want to mention the fact that this term was traditionally used interchangeably with the term "information retrieval", but as retrieval from other media (e.g., speech or images) has become more practical it is becoming more common to be explicit about the sort of information being retrieved.

2. 4 The Motivation for the Study of Cross-Language Text Retrieval

The recent enormous increase in the use of networked information access and on-line database has led to more databases being available in languages other than English. Oard and Dorr (1996) collect some examples, meant to be illustrative rather than exhaustive, which provide a further motivation for this research.

• A collection contains documents in such a large number of languages that it would be impractical to form a query in each language.

• The documents themselves are expressed in more than one language. For example:

- Technical documents in which English jargon appears intermixed with narrative text in another language.
- Literary criticism which quotes substantial portions of a work in a different text in another language.

• The user is not sufficiently fluent in a document collection's language to express a query in that language, but is able to make use of the documents that are identified. This would certainly be useful for a user who is able to read but not to write well in the document collection's language, but there are a wide variety of circumstances in which a reader, totally unfamiliar with the principal language of the document collection, might find CLTR useful. For example, a researcher seeking to determine which individuals and institutions have conducted research on a particular topic or a user with sufficient resources to translate the selected documents into a language that he or she is able to understand.

The authors conclude that a relationship between machine assisted translation and CLTR can be drawn from the last example. CLTR can be used to reduce the number of documents requiring translation, while machine assisted translation makes it practical to translate the selected documents at a reasonable cost. A similar relationship exists between CLTR and fully automatic machine translation. With such systems where fully automatic and machine assisted translation resources can be integrated with a CLTR system, queries can be constructed in whatever language they are expressed.

All these different kinds of situations make research in CLTR an attractive and interesting idea, and it is for this reason that I conducted my research work in this area of study.

2. 5 Fundamental Approaches to Cross-Language Text Retrieval

In this section I present the currently used approaches in CLTR. The taxonomy I follow is based on the work by Oard (1996, 1997b) and Fluhr (1995). Three main themes have emerged in the research literature: knowledge, corpus and text translation-based approaches. I will go through these approaches in the next subsections. Though MT techniques are being used recently for CLTR purposes, it is important to point out that this approach has not received as much attention as the other two.

2. 5. 1 Knowledge-based Approaches

In this section I present the different kinds of knowledge-based approaches in CLTR. Three main trends can thus be defined, thesaurus, dictionary and ontology-based. Most of the work I introduce in this section has to do with the research carried out using thesaurus-based approaches. This technique was pioneered in this field and consequently most of the work in CLTR has been centred on this approach.

Knowledge-based approaches are characterised by the vocabulary they can manage. Thus, at the narrower extreme, thesaurus-based approaches make use of what is known as "controlled vocabulary". A broader approach will allow the manipulation of a less restricted kind of terms and vocabulary. At this extreme, we find dictionary-based and ontology-based approaches. I will follow this classification, going from the narrowest to the broadest approaches.

2.5.1.1 Thesaurus-based systems

A thesaurus can be defined as an ontology that is specialised to organising terminology. A multilingual thesaurus is one which organises terminology from more than one language. Thesaurus-based techniques share some advantages. Because thesauri can represent relationships between terms and concepts in a way that humans find understandable, thesaurus-based text retrieval is conceived to allow users to reformulate better queries. Furthermore, because a significant amount of domain knowledge can be encoded in the thesaurus, in the hands of a skilled user a thesaurus-based text retrieval system can be a powerful tool. Nonetheless, the use of thesaurus-based techniques imposes certain limitations. Thus, thesauri impose an a priori limitation both on the vocabulary the user may employ and on the domain to which the text retrieval system can be applied.

Thesauri can be used either manually or automatically. In some systems known as "controlled vocabulary" systems, every concept is labelled with a unique descriptive term so that the user can manually specify the appropriate concepts in his or her query. When the concept relationships encoded in a thesaurus are used automatically, the technique is often referred to as "concept retrieval".

Controlled vocabulary CLTR systems are presently widely used in commercial and government applications for which the number of concepts (and hence the size of the indexing vocabulary) is manageable. Unfortunately, the requirement to manually index

the document collection makes controlled vocabulary text retrieval techniques unsuitable for high-volume applications in which the documents are generated from diverse sources that are not easily standardised.

The earliest reported experimental results on the effectiveness of CLTR were done by Salton at Cornell University in 1969 (Salton, 1970). Salton augmented his SMART text retrieval system with a multilingual concept list constructed by translating some of the words in an existing English concept list into German. He concluded that although retrieval effectiveness varied across document collections, "*cross-language processing ... is nearly as effective as processing within a single language*." After examining the retrieval failures in more detail he concluded that "*it would therefore seem essential that a more complete thesaurus be used under operational conditions for future experiments.*"

By 1973 it was well established that both controlled vocabulary and concept retrieval systems with multilingual thesauri could achieve performance across languages on a par with the within-language performance of the same techniques. Commercial acceptance followed soon, and by 1977 Iljon was able to identify four CLTR systems operating in Europe (Iljon, 1997).

The European Parliament's EUROVOC is an example of a modern multilingual thesaurus (Office for Official Publications of the European Communities, 1995). It includes all nine official languages of the European Community, and portions of it have been translated into additional languages. Thesaurus design remains expensive, and this fact has limited the domains to which controlled vocabulary retrieval has been applied. But EUROVOC demonstrates that once the basic concept relationships have been defined for a domain, extension of a multilingual thesaurus to additional languages is quite practical.

As large multilingual thesauri have proliferated, design and maintenance tools have become increasingly important. Thus, an automatic technique for using a thesaurus to

generate corresponding indexing terms in four languages was described by Pelissier and others in 1986 (Pelisser and Artur, 1986). In 1987 Kalachkina presented an algorithm for merging thesauri in different languages (Kalachkina, 1987) and in 1989 Loginov described tools developed in the Soviet Union to maintain a Russian-English version of the (monolingual) United States National Library of Medicine's Medical Subject Heading thesaurus (Loginov and V'yugin, 1989). Sogoaga of SABINI, a Spanish library automation company, also described the design of interactive tools for multilingual thesaurus maintenance (Sosoaga, 1991). The SABINI system was designed for automation of bibliographic records in an online library catalogue. Sogoaga provided no examples of implementations for specific languages, however.

More recently, a team at the University of Huddersfield Centre for Database Access Research in the United Kingdom led by Pollitt has integrated multilingual thesauri with interactive personal computer technology to address one of the fundamental limitations of controlled vocabulary text retrieval (Pollitt, Ellis et al., 1993). Experience has shown that although the domain knowledge that can be encoded in a thesaurus permits experienced users to form more precise queries, casual and intermittent users have difficulty exploiting the expressive power of a traditional query interface in exact match retrieval systems. Adapting their Menu-based User Search Engine (MenUSE) to use the European Parliament's multilingual EUROVOC thesaurus, Pollitt's team has developed a query formulation tool which facilitates visual browsing in the user's preferred language. The cited work does not report experimental results on the utility of the multilingual MenUSE interface.

CLTR systems are widely used today. Sophisticated multilingual thesauri have been developed for many domains and many languages, and the procedures for adding new domains and languages are well understood. In order to improve on present practice it is important to take into account the limitations of present systems. Thus, three main aspects that deserve special attention in this regard are cost, usability by untrained users, and effectiveness.

It is well known that thesaurus construction is an expensive activity. Although automated tools actually improve human productivity, as long as human intellectual activity is required to recognise and organise information the costs will remain substantial. In fact, human activities such as thesaurus maintenance and controlled vocabulary indexing have come to dominate system costs. This limits both the scalability of existing thesaurus-based systems to accommodate the rapid growth in electronically accessible texts and the generalizability of the technique to new domains for which construction and/or use of a thesaurus is economically impractical.

Another important limitation of controlled vocabulary text retrieval techniques is that untrained users seem to have difficulty exploiting their capabilities. In this sense, significant differences between the performance of skilled and untrained users have been observed with their choice of terms, their use of the term relationships that can be encoded in a thesaurus, and their use of operators such as *and*, *or* and *not* for query construction. In many cases, it has been proved that it is more economical to provide trained intermediaries than to provide adequate training to each user. Advanced users interfaces such as MenUSE offer some potential for mitigating this problem, and expert systems that construct Boolean queries from natural language have been investigated in a monolingual context (Marcus, 1994). The ranked retrieval techniques represent another approach to solving this problem. Ranked retrieval systems typically accept queries in natural language and allow a relatively unconstrained choice of terms.

Regarding effectiveness, it is important to consider the fact that language is a creative activity, and that new words enter human languages continuously. As thesaurus construction is time-consuming, thesauri in production applications necessarily lag behind the common use of terminology. Furthermore, there is some evidence that thesaurus designers have more difficulty anticipating which concepts and relationships will be useful to their system's eventual users than a cursory inspection of the thesaurus would suggest.