"A rose by any other name" Can you still find it? What you need to know about cross lingual retrieval

By Phil Vines and Sandra Potter

Phil Vines *is Deputy Head of the School of Computer Science and Information Technology at RMIT University. He obtained his PhD in 2001 in Chinese language information retrieval, and since that time he has continued to supervise PhD students and conduct research in cross lingual information retrieval. He is particularly interested in Asian language information retrieval, and has been active in the NTCIR workshop. He also holds a Bachelors degree in Chinese from the University of Melbourne.*

Sandra Potter *is a Director of Potter Farrelly Consulting and an internationally recognized expert in the law and technology field and is on the executive board of The Sedona Conference Working Group 6: International Electronic Information Management, Discovery and Disclosure*

Portions of this article are taken from an article prepared for the 2nd Annual Sedona Conference® International Programme on Cross-Border Discovery and Data Privacy held on 15-16 September in Washington, DC; they are reprinted courtesy of The Sedona Conference® (www.thesedonaconference.org).

WTO[1] initiatives and the growth of the global economy has made the world a small place when it comes to possible issues arising out of contractual agreements or information transfer. When Cross jurisdictional matters arise across borders the source language of such agreements and related communication becomes one of the most costly and challenging aspects of any case where the litigation is being run in one language and there are documents around the world in different countries in other languages. Automated cross lingual retrieval solutions have become an opportunity to cut some of the costs and still get the required results.

Background

It was not until the widespread usage of the World Wide Web in the Nineties that search engine usage became a common phenomenon. However research in to the best way to extract interesting relevant information from amongst large quantity of documents by automated means has been going on since the Fifties and Sixties[2]. With limited computational power, early systems often only indexed document titles and used Boolean keyword queries, rather than free text. Even today, the lay user may not understand that search engines still predominately use the bag of words model of information. Documents are simply seen as a bag of words without any significant regard to the structure or the meaning of the words. The search engine process works largely by matching words in a query with documents that contain those words. Try typing "Americans at war but not civil war" into your favorite search engines and chances are you will get the Wikipedia page on the American Civil War.

Yet despite these shortcomings, search engine technology has made enormous advances, particularly since the advent of the TREC[3] conference that started in the early nineties, and has probably been the most important catalyst in information retrieval research. The TREC conference quick developed a number of (ever changing) tasks known as tracks which concentrate in different areas of information retrieval. As well as simulating the behavior of a casual user of typing in queries and looking for relevant documents (the "ad hoc" task), there has been a wide variety of other tracks such as patent search, blog search, genomics track, question answering, and of course the legal track. The most important contribution of the conference is the provision of standardized document collections, query sets and relevance judgments, which provides a level playing field on which to compare the performance of different search engine software and this allow qualitative comparisons to be made, as well as providing a forum for researchers in the area to common together to share their experiences.

Most of the early search engine research was centered on English language documents and queries, as this was the dominant form of web content in the early period of the Web. As Web usage started to permeate non English speaking countries, the amount of non English content started to increase. One recent study estimated that slightly more than 50% of the total web content is now in Chinese. As the multilingual nature of the web started to develop, people started wondering if it would be possible to access the growing volume of information that existed in foreign languages. Some people dreamed of typing in a query in their own language, retrieving information in another language and then having it seamlessly automatically translated to their own language. While this Holy Grail is still some way off, substantial progress has been made.

In 1994, a Spanish language track was introduced at TREC, and in 1996 a Chinese track was added. European language based retrieval including combining English, French and German was later spun off into its own conference CLEF[4]. At that time, retrieval techniques for English and other related European languages were well advanced but less was known about the problems of indexing and retrieving information in other scripts. In 1999, the first NTCIR[5] conference was held in Japan. This conference looks at Chinese Japanese and Korean languages. Initially there were monolingual rather than cross lingual tasks, the idea being to give participants exposure to the problems of working in what was for most, a quite foreign language. In English language retrieval, documents are conceptualized as a series of words, and this is what is indexed. Chinese words consist of one or more characters, and thus there are substantial issues around what the unit of indexing should be. Further, with text being a continuous sequence of characters, with no spaces between words, word segmentation also becomes an issue. Attention quickly turned to cross lingual tasks, and the conference continues to the day. In 2001 TREC commenced and Arabic/English cross lingual track that ran for two years. A key part of cross lingual retrieval is of course translation. We give a brief overview in the major issues in the following section.

Language Translation Issues

Work on computerized language translation has been going on since the digital computers were built in the 1940's[6]. Researchers in the area have been quoted as saying that the problem should be solved in the next 10 years. Unfortunately they have been saying this for at least the last 30 years. While quality continues to improve, it falls short of what can be achieved with (costly) human translation.

Automatic translation generally proceeds on a number of levels, looking both at the translation of individual words and of the sentence structure. At the word level, the major issue is ambiguity. A word in a given source language may often have more than one meaning, and even if the meaning in a given context can be discerned, there may be multiple words in the target language that express different shades of similar meaning. A commonly cited example is bank, which may mean a river bank, a financial institution, or something else, such as in the plane banked sharply.

In standard language translation, there are a number of statistical techniques, which can be applied to select the correct translation based on the context in which a word appears in the source language. These typically used Markov based language models.

Language translation systems must analyze the grammatical structure of text to be translated. One of the major problems is that while there is generally a regular structure there are always exceptions and subtleties, which often confound machine translation systems. Humans can deal with these nuances, partly because in addition to the grammatical and syntactical clues, they also understand the meaning of what is being conveyed, which provides a good deal of extra information how a given structure should be understood. For example in the sentence Ross was told what to do by the river. The reader will dismiss the parse which means “the river told Ross what to do”, as it doesn't make sense, but an automatic translation system will not choose the correct structure every time, thereby leading to incorrect translations. It is generally acknowledged that some kind of representation of meaning is the key to high quality translation. This is the subject of much ongoing research.

Currently, good automatic translation systems can usually convey the gist of what is being said in the source language although there maybe issues with the structure of the translated material and the wrong choice of words. Greater success with languages that have somewhat similar structure, for example English to French, compared with English to Chinese, or English to Arabic. Aside from issues of difficulty of differing language pairs, the quality of a translation can vary greatly depending on the style of text being translated. Text that has long sentences with complicated structure (such as may be seen in legal documents) is much more prone to error, compared to short simple sentences. Text which attempts to convey meanings in subtle ways is similarly more prone to error that text which is straightforward and direct.

Language Translation in Information Retrieval

Cross language Information retrieval is generally defined as a situation where queries are in one language, and the “documents” being queried are in another language. Some people extend this to include the situation where documents are in multiple languages, although this is usually called multilingual information retrieval.

The first step in a cross lingual retrieval task is to convert it to a monolingual retrieval task and then apply standard monolingual retrieval approaches. This could be accomplished by translating the queries to the language of the documents, or the documents to the language of the queries. A third (less often used) approach is to translate both the documents and queries to a common third language. This might be a good strategy is situation where there may not be good translation resources available between a given language pair, e.g. Armenian and Japanese, but good translation resources exist between English and each of these languages.

Once the problem is reduced to one of monolingual retrieval, standard retrieval processes can be applied to retrieve documents that match the query. Normally, the documents will then need to be translated to query language (assuming the user is not familiar with document language) so that the documents can be evaluated. Past experiments in ad hoc cross lingual retrieval have shown that even though translations are not of high quality such that all meaning is clear and unambiguous, they are normally of sufficient quality to determine if the document is relevant to the query and thus worthy of more careful manual translation.

In the past the most common approach has been to translate the queries to the language of the documents, although recently there has been some interesting work on translating in both directions, performing two separate monolingual retrieval operations and merging the results.

The difficulty with translating documents to the language of the query is simply the standard language translation problem. It is hard and imperfect, as well as resource intensive. The problem with translating queries is the lack of context that is normally used to guide the selection of translation alternatives. However this is the same issue that occurs in monolingual retrieval. For example if a user simply types TREC into a search engine, without any context there is no way of knowing whether the user is interested in the Text Retrieval Experimental Collection, or the Texas Real Estate Commission.

Monolingual retrieval systems still largely follow the "bag of words" approach. That is, documents are retrieved on the basis that they have words in common with the words in a query. Documents with more words in common rank more highly, and having relatively rare words in common is more important than having frequently used words in common. In particular, the syntactic structure of the query is of little importance for effective query translation. Because of this, translation of simple queries can be quite effective if the meaning of the words is translated correctly, even if the syntax and grammar of the query are not. For this reason it has been most popular choice.

Query Translation

There are two major challenges in query translation: Translation Ambiguity, and Out Of Vocabulary (OOV) words. The problem of ambiguity is similar to that which occurs in standard translation, namely that many words have more than one meaning. A translation process might typically commence by looking up each word in a translation dictionary and finding the possible alternative translations. To take a concrete example, suppose a query was "bank interest rate increases". If the translations were found for each of these words, we might find a number of alternatives for each word. Some method is needed to select the most appropriate translation.

Fortunately we can exploit language patterns to solve this problem. Provided we have a large amount of text available in the target language, we can observe patterns of word associations. The translated words for bank (financial institution) and interest (money) will tend to occur near each other far more often than other alternative combinations such as bank (financial institution) and interest (interested in), or bank (river) and interest (money). By employing appropriate statistical techniques we can compute the statistically most likely translation.

Most of the time this approach works well, provided there is enough contextual information in a query. If however the query itself leaves room for ambiguity, for example "java books", which could mean the island, the programming language, or coffee beans, then a single translation may well miss the meaning completely, although it should be noted that the same problem may arise in a monolingual search engine as well.

Rather than having to decide on the single most likely translation, some search engines, especially research vehicles such as Indri[7], allow the translation engines the luxury of hedging their bets, and select more than one translation for each word, and attach probabilities to the translations. This tends to increase coverage, i.e. the proportion of the total number of relevant documents found, but decreases precision (the proportion of documents found that are relevant).

Out of Vocabulary words and phrases

Most query translation systems use some kind of a dictionary for part of the process. If the word to be translated is not found in the dictionary this may cause a major problem. This issue will most often occur with so called named entities which include things such as personal names, company names, place names, book and movie titles. When these terms occur in a query they are often the key term and failure to translate this term correctly can be disastrous. This is more likely to be the case with highly dissimilar language pairs, for example English and Chinese, or English and Arabic. For more similar language pairs, such as English and French, simply using the foreign language word direct in the query without translation, has been shown to be quite effective. The reason for this is that in such languages the foreign terms are often used directly, without translation, in the other language.

In languages where OOV words need to be translated, various techniques that have been developed to minimize this problem. Once an OOV word is identified (one not in the translation dictionary), search engines can be employed to search the entire web for this term. In some language pairs, for example English and Chinese, it is common that when a new Chinese word is used, being a translation of a foreign term, then the foreign term may sometimes be used nearby. By applying appropriate techniques, the translation can be inferred, with high probability.

It is important to understand that such techniques will work most if the time, but not all the time. For example, a new word may be coined in one language to name a new product or to describe a newly invented process. When the word is first coined, there is almost certainly not going to be translations of this term into other languages, no matter how good your system is. Sometimes what then happens is that initially a number of competing translations emerge, and then after a period of time one becomes more dominant.

Phrases

A further challenge in translation is the recognition and translation of phrases, or more generally, two or more words that have a special meaning when they occur together that is different from the translation of the individual words, for example acid test or cloud nine, or special legal phrases, such as due diligence, or intellectual property or duty of care. If these are not recognized as phrases at translation time but instead translated as individual words, the result will inevitably be incorrect. All commercial translation systems will have a phrase detection component; however they will differ in the range of phrases that they recognise.

Language specific issues

Many languages have specific issues associated with them that need to be catered for. For example, the Japanese language makes use of several different scripts including kanji (derived from Chinese) and hiragana and katakana (phonetic scripts). Some words can be written either in kanji or a phonetic script. Without special precautions, searching for a word in one script will not locate a document that uses it in the other script. The coding schemes employed may also be an issue. Unlikc English documents that are coded almost exclusively in ASCII, many foreign languages have more than one coding "standard". Mainland China uses gb2122 and simplified characters, whereas other Chinese speaking regions tend to use "Big-5" and traditional characters.

Evaluation

The standard evaluation process involves a test collection and a set of queries. For each query a search engine is required to retrieve a large number of documents, for example, the top 1000 documents. The more relevant documents a search engine finds, the better it is, and the more highly ranked these documents are the better the score. The metric which is used to characterize this is called Mean Average Precision or MAP. For example finding some relevant documents and ranking them in the first hundred is far better than finding the same relevant documents and ranking them in the last. This mimics the preference of the casual search engine user who is normally happy to find just one, or at most a small number of relevant documents near the top of the list of documents. It effectively favours high precision, that is, finding relevant documents early, over recall, which is the proportion of the total number relevant documents that were retrieved.

When reporting the results of cross lingual retrieval experiments, it is common to report results as a percentage of monolingual performance. So one might read a claim that a particular search engine achieved a MAP which was 90% as good as the Monolingual result. While in one sense this is impressive, it needs to be remembered that most monolingual tests usually only retrieve a proportion of the relevant documents. Sometimes a cross lingual system's performance may even exceed 100% of the monolingual baseline. Usually what is happening here is a form of query expansion. Sometimes when a word has translations that are different but nonetheless similar in meaning the query will turn up other relevant documents that had words which were similar but not identical in meaning to the original terms. For example bank may be translated to words that mean finance corporation and building society in the foreign language and this might help to locate additional relevant documents.

Challenges for Cross lingual legal retrieval

From the above it can be seen that while cross lingual retrieval is quite feasible, there are plenty of pitfalls for the unwary. The important thing is to be aware of such issues and understand how a given Cross Lingual systems works. What does a system do if it cannot translate a word? Does the system only produce one translation for each word, or several? Most translation systems can handle common phrases, but not necessarily more unusual or specialist (legal) phrases. If it is the queries that are being translated from English to a foreign language then a simple inspection of the query will give some indication of the potential risks. Are there any ambiguous terms, are there any proprietary names, or special phrases?

Most research and development in search engines has generally concentrated on finding the best documents in response to a query. Legal retrieval is somewhat different in that we are trying, as far as practicable to find all the documents that are relevant to a query. While the ranking of documents in response to a query is not directly useful in a situation where we are trying to find all relevant documents, it nonetheless gives us some clue as to how far down the list we should look. If we evaluate documents in the rank order that they are returned, we will get more relevant documents near the top of the list and a decreasing proportion of relevant documents the further we go down the list. When the proportion of relevant documents being found becomes sufficiently low, using some form of cost benefit criteria, we can stop.

[1] WTO – World Trade Organization

[2] Cyril Cleverdon, Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, *The College of Aeronautics, Cranfield*, 1960

[3] http://trec.nist.gov

[4] http://clef2010.org/

[5] http://research.nii.ac.jp

[6] From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947–1954. A Chronology, Journal of Machine Translation, (springer) Vol 12, no. 3, 1997. pages 195-252

[7] Indri: a language-model based search engine for complex queries by Trevor Strohman, Donald Metzler, Howard Turtle, W. Bruce Croft — 2005 — in Proceedings of the International Conference on Intelligent Analysis