

Literature Review of Cross Language Information Retrieval

Mustafa Abusalah John Tait Michael Oakes

Abstract-Classical Information Retrieval (IR) is the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents. A search engine performs IR by retrieving relevant web pages from the internet. Rather than regarding foreign-language documents simply as unwanted "noise", Cross Language Information Retrieval allows the user to state their query in one language, and retrieve documents in another. Some CLIR systems use language resources such as bilingual dictionaries to translate the user's original query, while other systems use machine translation to translate the foreign-language documents beforehand, enabling them to be retrieved by the original query. Problems arise due to ambiguity in language, the use of synonyms to express a single idea, and the lack of context available in translating a short query. This paper will discuss previous work in CLIR, current problems in CLIR, and make recommendations for future work.

Keywords-Cross Language Information Retrieval, Lexical Semantics, Disambiguation, Translation.

1. INTRODUCTION

In classical IR, both the query and the documents are in the same language. The basic idea behind the cross language information retrieval (CLIR) system is to retrieve documents in a language different from a query language made in the user's own language. This may be desirable even when the user is not a speaker of the language used in the retrieved documents. Once it is known that the information exists and is relevant, the retrieved documents can be translated by a human translator. For example, when doing original research, it is essential to find out whether the topic of interest has already been studied elsewhere in the world.

This report will focus on current approaches to CLIR systems. In particular we will consider machine-translated queries, dictionary-based query translation, and the use of parallel corpora.

2. CLIR APPROACHES

2.1 Machine Translation Approach

In CLIR, Machine Translation (MT) can be implemented in two different ways. The first way is to use an MT system to translate foreign language documents in the corpora into the language of the user's query. This is done off-line beforehand. This approach is not viable for large document collections, or for collections in which the documents are in numerous languages. For example, in his experiments on German-Spanish CLIR, (Braschler et al 2004 [1]) was not able to find direct German/Spanish MT so he had to use German/English MT, then English/Spanish

MT. Not all the terms in the original German documents could be translated by this "triangulation" process.

In the second method of using MT in CLIR, the users query in the "source" language is translated into the "target" language (the language of the documents in the stored collection). The "target" language query is then used to retrieve "target" language documents using classical IR techniques.

With both methods, the MT stage is separate from the retrieval stage. An ambiguity problem exists in the MT component, since the translated query does not necessarily represent the sense of the original query. For instance, translating the English query *big bank* to another language could produce an inappropriate translation since it is not clear whether "bank" means the institution or the edge of a river. MT systems normally attempt to determine the correct word sense for translation by using context analysis (Braschler et al 2004 [1]). However, a typical search engine query lacks context as it consists of a small number of keywords. MT is more efficient in documents translation as the context is clearer.

2.2 Dictionary-based query translation Approach

In dictionary based query translation the query keywords are translated to the target language using Machine Readable Dictionaries (MRD). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination of both. The major problem in the bilingual dictionary approach is translation ambiguity (as is the case for MT systems, discussed in Section 2.1) in addition to problems of word inflection, problems of translating word compounds, phrases, proper names, spelling variants and special terms (Ballesteros and Croft 1997 [5], 1998 [6], Hedlund et al 2004 [8]).

2.3 Translation Disambiguation

Two of the causes of ambiguity in natural languages are homonymy and polysemy, where homonymy refers to a word that has at least two entirely different meanings (such as "bark", which can mean the skin of a tree or the voice of a dog) while polysemy refers to a word which can take on two distinct, but related meanings (such as the "head" of the body, and the "head" of a department). The distinction between homonymy and polysemy is not always clear cut (Akmajian et al. 1990 [15], Lyons 1984 [16], Kilgarriff 1993 [17]). Lexical ambiguity covers both homonymy and polysemy. Translation ambiguity arises due to source and target language lexical ambiguity. Many methods have been developed to decrease the ambiguity in query translation, such as part of speech tagging, corpus based disambiguation methods, query structuring (Ari Pirkola et al 2001 [13]), and the most probable translation strategy (Kraaij 2001 [3]), as described below:

1. In Part of speech tagging, words are annotated to their part of speech (noun, verb, etc.). This can be done using the Xelda toolkit (Kraaij, 2001 [3]).
2. The Corpus-based disambiguation technique involves query expansion (Q) to reduce the effects of bad translation equivalents (Ballesteros and Croft, 1996 [4], Ballesteros and Croft 1997 [5], Chen et al. 1999).
3. Query structuring can be applied in different situations such as Syn-based structuring, Compound-based structuring, phrase-based structuring (Ari Pirkola et al 2001 [13]).
4. By using the most probable translation strategy, a single translation for each query term is selected based on the number of occurrences of translations in the dictionary. For example, the English word *bar* has more than one identical translation for different senses. The French words for “bar” could be “barreau” (window bar), “barreau” (legal bar), “tablette” (bar of chocolate). If these were the only three senses of “bar” in the dictionary, the “most probable” translation would be “barreau”, since this is the translation for 2 out of the 3 senses. The implicit assumption in this strategy is that the number of occurrences of a translation in the dictionary may serve as a rough estimate of an actual translation probability. Ideally, these probabilities should be estimated from actual corpus data (Kraaij 2001 [3]).

Short queries are usually insufficient to describe the need of the user in a precise and unambiguous way, and this makes the above steps harder and sometimes insufficient. For example, a query with the keywords **Turkey Arms**, could refer to either the Turkish Army, or the bird turkey wings.

2.2.1 Word inflection

A common problem with query translation is word inflection. This can be solved by lemmatization, where every word is reduced to its uninflected form or lemma. Another technique is called stemming, where different grammatical forms of a word are reduced to a common shorter form (not necessarily the lemma) called a stem, by the successive removal of word endings. For example, the stemming rules “remove -ion”, “remove -at”, and “remove -ity” will transform both **Gravitation** and **Gravity** to “Grav-“ (Belew 2001 [23]).

2.2.2 Phrases

For the success of CLIR, translation of phrases in their entirety, rather than individual word-for-word translation, is crucial. (Hull and Grefenstette 1996). Phrases matched against a manually built multi-word (phrase) dictionary showed higher precision than those translated by single word-based dictionaries (Ari Pirkola et al 2001 [13]).

2.2.3 Compound words

A compound word is a word formed from two or more words; compound words are not widely available in English, but very much used in other languages such as German, Finnish, etc. A compound word can be decomposed to two or more words, where each has a meaning are called compositional compounds, for example a Finnish word *kaupunginhallitus* (*city government*) is decomposed into two components, each of which has a meaning, *kaupungin* (*city*) and *hallitus* (*government*); but the problem occurs with non-

decomposable compounds whose meaning can't be deduced on the basis of its components, or semi-compositional compounds with meanings that in part could have meaning but not related to the full compound meaning, for example the Finnish compound *krokotiilinkyneleet* (*crocodile tears*). Compound splitting can be performed effectively by means of a lexicon-based morphological analyzer. (Ari Pirkola et al. 2001 [13]).

2.2.3 Proper names and spelling variants

In many documents technical terms and proper names are important text elements. Their translation is crucial for a good CLIR system (Ari Pirkola et al. 2003 [9]). MT lexicons and general bilingual dictionaries lack translations for proper names and spelling variants. A common method used to handle untranslatable keywords is to include them untranslated in the target language query. If this word does not exist in the target language, the query will be less likely to retrieve relevant documents. Alternative methods exist to solve this problem for languages of the same writing system such as Transformation rule based translation (TRT). In TRT a word in one language is matched to a word in other language based on regular correspondences between the characters of the two languages. Thus the source language vocabulary and the target language vocabulary are regarded as spelling variants of each other. For example, the Spanish word “embriologia” matches “embryology” in English by replacing -ia with -y. (Ari Pirkola et al. 2003 [9]). Usually TRT is used in conjunction with fuzzy matching such as the n-gram matching technique. In the n-gram method search keywords are decomposed into n-grams (sub-strings of length n), then the degree of similarity is computed by comparing their n-gram sets (Pfeifer et al. 1996 [18], Robertson and Willett 1998 [19], Žebel and Dart 1995 [20]).

2.2.4 Special terms

Special terms are most likely to be technical or scientific terms that are not widely available in general dictionaries. Special terms can be matched against a special dictionary, e.g. a medical term can be matched against a medical dictionary. Combining both general and specific domain dictionaries enhances the retrieval results. Two techniques are used to combine both dictionaries. Sequential translation translates the query keywords against the specific domain dictionary. If it fails to match, it uses the general dictionary, and a parallel translation that matches query keywords against both general and specific dictionaries. Both these techniques reduce the special terms translation problem but don't solve it altogether. For instance, translating a newspaper article that contains scientific terms, technical terms, political terms etc. needs more than a domain specific dictionary.

2.4 Corpus-based Approach

A Corpus is a repository of a collection of natural language material, such as text, paragraphs, and sentences from one or many languages. Two types of corpora (plural of “corpus”) have been used in query translation:

2.4.1 Parallel Corpora

Parallel corpora consist of the same text in more than one language. An aligned parallel corpus is annotated to show exactly which sentence of the source language corresponds

with exactly which sentence of the target text. When retrieving text from a parallel corpus, the query in this does not need to be translated, since a source language query can be matched against the source language component of the corpus, and then the target language component aligned to it can be easily retrieved. Parallel corpora can be populated using human translation, websites in more than one language or using MT methods. "Spider" systems have been developed to collect documents that have translation equivalents over the internet to produce corpora.

The alignment process can be done by comparing documents by the presence of indicators. The indicator could be an author name, document date, source, special names in the document, numbers or acronyms, in fact anything which clearly corresponds in both the source and target language texts. Another example of parallel corpora alignment is the PTMiner tool (Nie, Simard, Isabelle and Durand 1999 [14]). The system first determines candidate sites, and then identifies a set of web pages on each web site that are indexed by a search engine. The next step is to construct pairs of web pages on the bases of pattern matching between URLs (index.html vs. indexf.html). The final step is to filter the candidate parallel page. (Kraaij et al. 2003 [7]). Yet another alignment method was developed for bilingual reports of election results (Braschler and Schäuble 1998 [12]).

2.4.2 Comparable Corpora

Comparable corpora contain text in more than one language. The texts in each language are not translations of each other, but cover the same topic area, and hence contain an equivalent vocabulary. A number of statistical techniques can be used to derive topic-specific (often technical) bilingual dictionaries from parallel corpora.

3. CONCLUSIONS

CLIR systems, approaches, and implementations are not limited to the discussion above as CLIR is one of the hot topics in the field of IR. Currently the best known search engines over the internet (excluding Google) use monolingual search (classical IR) only as CLIR systems are not generally available. CLIR systems that have combined query translation to parallel-corpora could show better results as parallel corpora has rich context to cover the weak context of the query.

The semantic web can play an important role in CLIR through the use of ontologies. An *ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary* (Neches et al., 1991 [21]). An ontology is an explicit specification of a conceptualization (Gruber, 1993 [22]). Ontologies can be implemented in translation systems to extract conceptual relations for monolingual and cross language IR.

4. REFERENCES

- [1] MARTIN BRASCHLER:Combination Approaches for Multilingual Text Retrieval *Eurospider Information Technology AG, Schaffhauserstrasse 18, CH-8006 Zurich, Switzerland; Universit e de Neuch atel, Institut Interfacultaire d'Informatique, Pierre- a-Mazel 7, CH-2001 Neuch atel, Switzerland.* Information Retrieval, 7, 183204, 2004
- [2] Ari Pirkola & Turid Hedlund & Heikki Keskustalo & Kalervo Jarvelin: Cross-Lingual Information Retrieval Problems: Methods and findings for three language pairs. *ProLISSa Progress in Library and Information Science in Southern Africa.* First biannual DISSAnet Conference. Pretoria, 26-27 October 2000.
- [3] Wessel Kraaij and Ren e Pohlmann. Different Approaches to Cross Language Information Retrieval. In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zvrel, editors, *Computational Linguistics in the Netherlands 2000*, number 37 of Language and Computers: Studies in Practical Linguistics, pages 97-111. Rodopi, Amsterdam, 2001.
- [4] Ballesteros, and Croft, B. Dictionary Methods for Cross-Lingual Information Retrieval. 7 th DEXA Conf. on Database and Expert Systems Applications. Pages 791-801, 1996.
- [5] Ballesteros, L., and Croft, B. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. Pages 84-91, SIGIR 1997.
- [6] Ballesteros, L., and Croft, B. Resolving Ambiguity for Cross-Language Retrieval. Pages 64-71, SIGIR 1998.
- [7] Wessel Kraaij, Jian-Yun Nie, and Michel Simard. Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval. *Computational Linguistics*, 29(3):381-419, 2003.
- [8] TURID HEDLUND, EIJA AIRIO, HEIKKI KESKUSTALO, RAIJA LEHTOKANGAS, ARI PIRKOLA, KALERVO J  ARVELIN: Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000/2002 *Department of Information Studies, University of Tampere, Finland Received December 5, 2002; Revised May 14, 2003; Accepted May 14, 2003* Information Retrieval, 7, 991-19, 2004
- [9] Ari Pirkola, Jarmo Toivonen, Heikki Keskustalo, Kari Visala, Kalervo Jarvelin: Fuzzy Translation of Cross-Lingual Spelling Variants 2003 ACM 1-58113-646-3/03/0007
- [10] Craig J.A. McEwan , Iadh Ounis and Ian Ruthven Building Bilingual Dictionaries From Parallel Web Documents European Conference on Information Retrieval (ECIR 2002), Glasgow, March 2002
- [11] Mark W. Davis and William C. Ogden: Free Resources And Advanced Alignment For Cross-Language Text Retrieval, NIST Special Publication 500-240:The Sixth Text Retrieval Conference (TREC 6) page 385.
- [12] Braschler M and Sch uble P (1998) Multilingual information retrieval based on document alignment techniques. In: Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL '98, Lecture Notes in Computer Science, Vol. 1513, Springer, pp. 183-197.
- [13] Ari Pirkola, Turid Hedlund, Heikki Keskustalo, Kalervo J avelin: Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Inf. Retr.* 4(3-4): 209-230 (2001).
- [14] Jian-Yun Nie, Michel Simard, Pierre Isabelle, Richard Durand: Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. SIGIR 1999: 74-81
- [15] Akmajian, A., Demers, R., Farmer, A., & Jamish, R. (1990). *Linguistics: An Introduction to Language and Communication.* Chapter 2: Morphology: The Study of The structure of Words. (pp.11-52). Cambridge: MIT Press.
- [16] Lyons, J. (1984), *Language and Linguistics: An Introduction*, Cambridge: Cambridge University Press.
- [17] Adam Kilgarriff, *Dictionary Word Sense Distinctions: An enquiry into their nature*, *Computers and the Humanities* 26 (1-2), pp 365-387, 1993.
- [18] Pfeifer U, Poersch T and Fuhr N (1996) Retrieval effectiveness of proper name search methods. *Information Processing & Management*, 32:667-679.
- [19] Robertson AM and Willett P (1998) Applications of n-grams in textual information systems. *Journal of Documentation*, 54(1):48-69.
- [20] Z ebel J and Dart P (1995) Finding approximate matches in large lexicons. *Software practice and experience*, 25(3):331-345.
- [21] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. Swartout. *Enabling Technology for Knowledge Sharing.* *AI Magazine*, 12(3):37--56, 1991.
- [22] Gruber, T. R. "A translation approach to portable ontology specifications". *Knowledge Acquisition*. Vol. 5. 1993.
- [23] R.K. Belew. *Finding Out About* [Cambridge Univ. Press, 2001] ISBN 0-521-63028-2

Mustafa Abusalah: PhD Computer Science student at University of Sunderland, UK, and computer science instructor at the Arab American University, Jinen, Palestine, email: Mustafa.abusalah@sunderland.ac.uk

John Tait: professor of Intelligent Information Systems, University of Sunderland, UK, email: john.tait@sunderland.ac.uk

Michael Oakes: Phd, Senior Lecturer, University of Sunderland, UK, email: Michael.Oakes@sunderland.ac.uk