Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base

Mustafa Abusalah, John Tait, and Michael Oakes

Abstract—**This paper reports an experiment to evaluate a Cross Language Information Retrieval (CLIR) system that uses a multilingual ontology to improve query translation in the travel domain. The ontology-based approach significantly outperformed the Machine Readable Dictionary translation baseline using Mean Average Precision as a metric in a user-centered experiment.**

***Index terms*—Ontology, multilingual, cross language information retrieval.**

I. Introduction

THE growing requirement on the Internet for users to access information expressed in language other than their own has led to Cross Language Information Retrieval (CLIR) becoming established as a major topic in IR. One approach to CLIR uses different translation approaches to translate queries to documents and indexes in other languages. As queries submitted to search engines suffer lack of context, translation approaches have great problems with resolving query ambiguity. In our approach, we built a multilingual ontology to be used as a translation base for CLIR. In this paper we evaluate our proposed query translation methodology and compare it with a base line system that uses a Machine Readable Dictionary (MRD) as translation base in a user-centered experiment.

II. Background

CLIR approaches are decomposed into two research fields, the first is bilingual MRD and machine translation (MT), and the second is concept driven approaches.

The major problem in the bilingual dictionary approach is translation ambiguity in addition to problems of word inflection, problems of translating word compounds, phrases, proper names, spelling variants and special terms [8], [9], [10]. MT systems normally attempt to determine the correct word sense for translation by using context analysis [11]. However, a typical search engine query lacks context as it consists of a small number of keywords. MT is more efficient in document translation as the context is clearer.

Concept driven approaches such as thesauri and multilingual ontologies bridge the gap between the linguistic term and its meaning.

A Bilingual Thesaurus groups words with similar meanings in hierarchies (with several levels) of classes and sections and maps them according to their meanings. EuroWordNet is an example of a multilingual thesaurus that uses "is-a" relations (amongst other types of relations) between "synsets", or groups of synonymous words and maps them according to their meanings using a bilingual index. However, the thesaurus does not include the definition of words. In fact, words in a group are merely related, not synonymous. In addition, words under a common heading can be of different syntactic categories. EuroWordNet groups terms of synsets with basic semantic relations between them.

In our approach we considered developing a bilingual ontology rather than collecting a thesaurus, because we consider ontology as a generalized collection of knowledge that will be used to add a context to search queries by the query expansion, enabling word sense disambiguation. Ontology defines concepts, terms and vocabulary in a domain, and also the relationship among these concepts. Concepts are organized in a taxonomic structure, with subclasses inheriting properties and specializing from superclasses. Current semantic web technologies also have the added capability of inferring new facts from old facts already captured in the ontology. An ontology, together with a set of instances of the classes or concepts defined, constitutes a knowledge base about the domain being described [12].

III. Ontology versus MRD

The ontology was built to model the travel domain and decomposed into two ontologies (Arabic and English Ontologies). The ontology was developed manually with the help from a domain expert. Both ontologies are mapped using an English Arabic bilingual index. The manually created ontology consists of 100 English concepts mapped to their Arabic equivalents and it was updated with 100 English concepts mapped automatically to the equivalent Arabic concepts a total of 200 mapped concepts. The automatic ontology mapping process that applied WSD (Word Sense Disambiguation) scored a precision of 0.83 in a user based evaluation. In addition to concept relations, such as "is a" and "has a" relationships, ontology also includes "instance of" and many other relations. Those relationships are represented in ontology languages like owl and rdf constructs. Concept

Manuscript received October 14, 2008. Manuscript accepted for publication August 3, 2009.

The authors are with the School of Computing and Technology, University of Sunderland, Sunderland SR6 0DD, UK, (e-mail: mustafa.abusalah@sunderland.ac.uk, john.tait@sunderland.ac.uk, michael.oakes @sunderland.ac.uk).

relations are used to expand queries with semantically related concepts to improve the information retrieval system's monolingual and cross lingual effectiveness. For example "Hotel" is a sort of "Accommodation", so if "hotel" was a query keyword it will be expanded to hotel or accommodation to return more relevant results in monolingual retrieval and referred to its equivalent concepts in Arabic to return more relevant results in Cross Lingual retrieval. In the retrieval system the ontology is combined with an MRD so if the ontology did not succeed in translating concepts, the MRD will translate them, and the translated query will be a combined translation of the ontology and the MRD. The ontology was constructed prior to the experimental query set being identified. It was developed using Protégé as it allows the developers to create, browse and edit domain ontologies in a frame-based representation. In addition plug-ins to enhance ontology development such as the OWL plug-in, were used to develop the OWL ontology. Both ontologies, Arabic and English, are mapped at the semantic level; each concept in both ontologies is mapped to its equivalent concept using a bilingual index defined in the English Ontology. We have developed an automatic ontology mapping tool to define and execute semantic bridges to map both ontologies. Figure 1 demonstrates a simple ontology translation process.

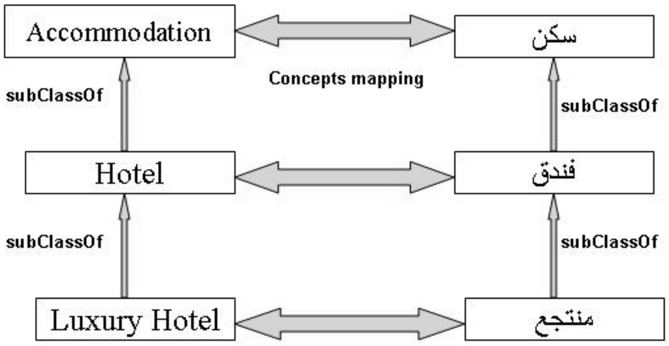

Fig. 1. Simple Concept Matching Task.

As a base for our Information Retrieval system we used the full text search technique. Full text search (also called free search text) refers to a technique for searching corpora; in a full text search, the search engine examines all of the words in every stored document as it tries to match search words supplied by the user. Some Web search engines, such as AltaVista, employ full text search techniques.

In our approach to employ full text search we generated a complete index for all of the searchable documents in the corpora. For each word (excepting stop words which are too common to be useful) an entry is made which lists the exact position of every occurrence of it within the database of documents. From such a list it is relatively simple to retrieve all the documents that match a query.

The MRD is Al-Mawred English Arabic dictionary [1] which has 100,000 English/Arabic entries and 67,000 Arabic/English entries. As noted above, it is used for MRD based CLIR as a baseline and to augment the ontology based translation. The Dictionary based IR system passes each query keyword to the Arabic/English Dictionary and the results are submitted to the search engine. In the dictionary model when a keyword is translated and has many synonyms the first matched synonym is selected. Figure 2 shows CLIR using MRD.

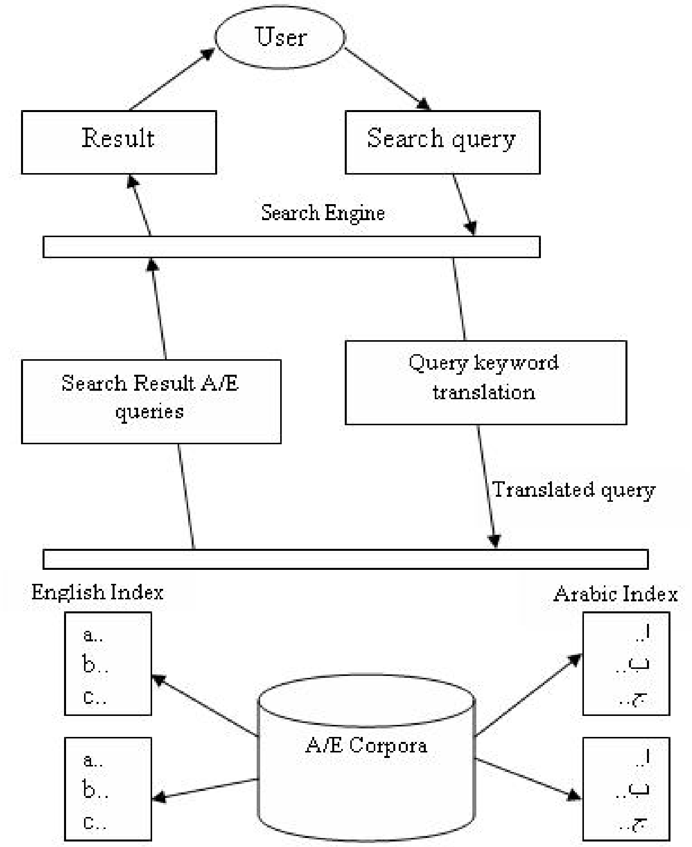

Fig. 2. Shows the CLIR process using MRD.

The Ontology based IR system submits the query keywords to XSL (Extensible style sheet Language) to query the ontologies, extracting related concepts and concept relations. Then concepts associated with semantic relations are studied by the ontology based CLIR system and identified for query expansion if synonyms were found, this is all done monolingual, then concepts are translated into their equivalent concepts in the other language using the ontology bilingual index. If the concept was not found in the ontology, the Dictionary is used to find the relevant translated concepts. Figure 3 shows the CLIR process using ontologies. In both dictionary and ontology based CLIR systems the final translated query terms are combined using the Boolean OR and then matched with the corpora documents. The results then are ranked depending on many factors such as the number of matching terms found in each document and the number of terms occurring in the document. We used the BM25 [13] (Best Match) weighting scheme to rank the found documents. TREC tests have shown BM25 to be the best of the known probabilistic weighting schemes [14].

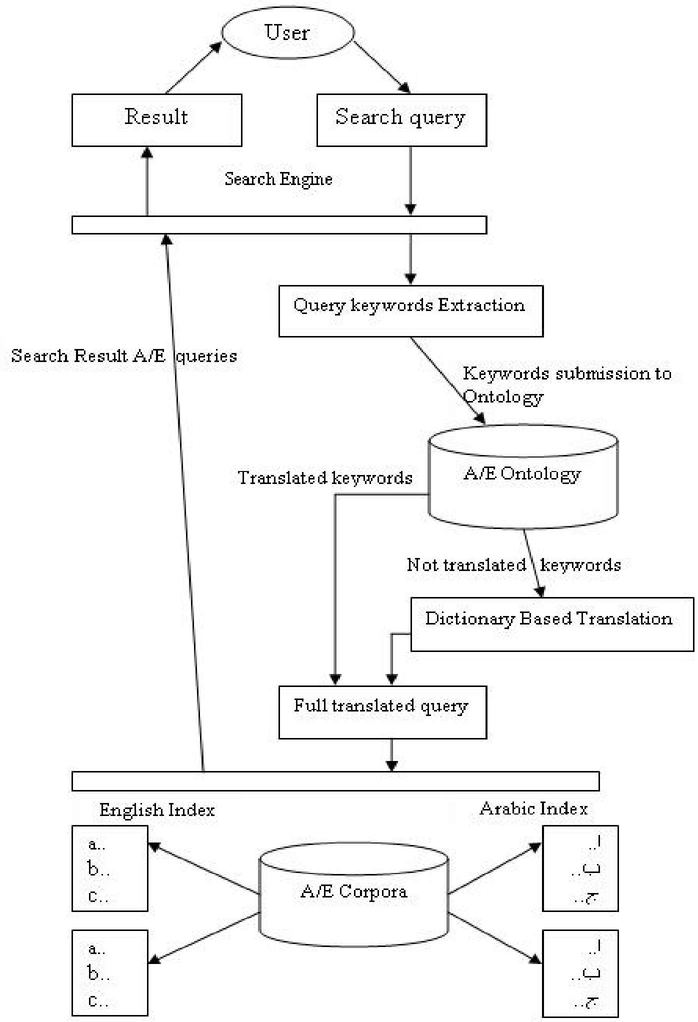

Fig. 3. Shows the CLIR process using ontology.

IV. Evaluations

The evaluation is based on human relevance assessments during experimental search sessions. 25 common queries were identified by discussion with experts in the travel and tourism field as being of interest to potential users of a travel search engine. They were expressed in the English language. The judges who evaluated both systems are not experts in the travel field but they have at least traveled abroad once. All judges are native Arabic speakers and have a very good knowledge of the English language. Twenty judges made a relevance judgment for each query submitted to each system with a total of 1000 judgments for the 25 queries for both systems. The judges access the system using a web-based interface, and submit the queries to both systems. We conducted two experimental runs.

Run 1: The Judge submits the query to the dictionary based system and evaluates the first 40 results appearing on the web browser with title and brief description.

Run 2: The same procedure applied in run 1 is applied in run 2 but using ontology based translation.

The judgment was binary as to whether result was relevant or not [2]. The judges were asked to score the quality of relevance match according to one of the following relevancy scale (not relevant, don’t know, possibly relevant, relevant, critically relevant), as shown in figure 4.

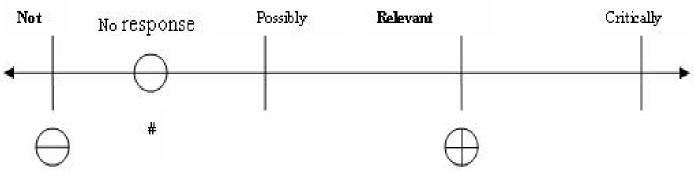

Fig. 4. Relevance Scale.

These responses were mapped onto a binary scale relevant if the document retrieved is at least possibly relevant, otherwise the document is not relevant. For example, critically relevant documents specifies to the user exactly what he is looking for, while possibly relevant might have some useful information, but doesn't specify exactly the user need. If the judge did not know whether the document is relevant or not his judgment is considered not relevant [7]. The document collection used in this experiment is about 8,000 documents in the Arabic language. The documents are all related to the travel domain and either published in Al-Nahar newspaper [3] from the year 1996-1999 or documents collected from the Palestinian ministry of tourism [4]. The scale of the collection, together with only two related systems being used in the experiment, meant no reliable assessment of recall could be made within the available time and resources.

V. Results

After the users’ assessment the Mean Average Precision [5] was measured, where Mean Average Precision is the average of the precision after each relevant document is retrieved.

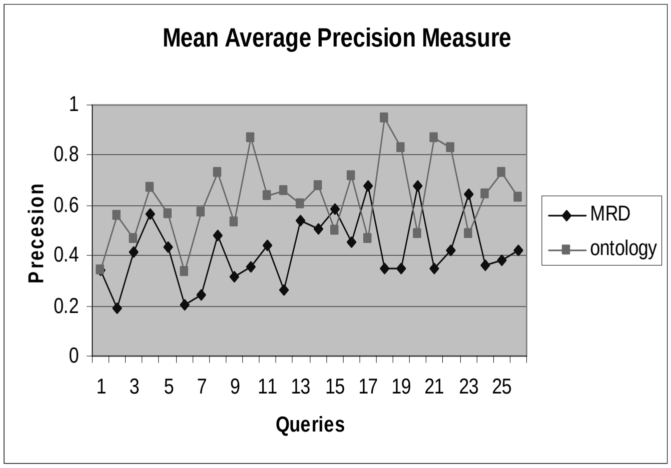

Fig. 5. Ontology versus Dictionary precision measure result.

$$MAP = \frac{\sum_{r=1}^{N}(pr(r)\times rel(r))}{Number\ ofRelevant\ \ Documents} \qquad (1)$$

where *r* is the rank, *N* is the number retrieved, *rel()* is a binary function on the relevance of a given rank, and *Pr()* is precision at a given cut-off rank.

Figure 5 shows the measurement of mean average precision for both the Dictionary and ontology based CLIR systems. The first run that measured the dictionary based CLIR system scored average MAP result **0.42** while run two that measured the ontology based CLIR system scored average MAP result **0.63** which is much better than the Dictionary based system average MAP result.

VI. Conclusions and Discussion

In this experiment, the effectiveness of the ontology based CLIR was better than the Dictionary based one. The benefit of using ontology is not limited to normal word to word translation. These results are especially interesting because they contrast with early monolingual work (e.g. Voorhees [6]) in which this sort of query expansion degraded rather than improved retrieval effectiveness. It is difficult to determine at this stage whether the improvement is a product of operating in a narrow (and known) domain, the scale and variety of the document collection or some other cause.

After the evaluation of both the pure dictionary and the ontology systems, the ontology based system scored higher in terms of precision. In future development we will enhance and extend the ontology by using annotation tools to align new concepts to the ontology and then test it again with the dictionary system. Other areas for investigation include ease of use, the use of relevance feedback, the effect of more extensive use of concept relations and possibly experiments with larger data sets.

References

[1] D. A. Lelmalayin, http://www.malayin.com/index_e.asp 26-01-2006.

[2] F. C. Gey, “The TEC-2001: Cross Language Information Retrieval Track,” 2001.

[3] Evaluations and Language Resources Distribution Agency, http://www.elda.org 26-01-2006.

[4] Palestinian ministry of tourism, http://www.visit-palestine.com/ 26-01-2006.

[5] J. Levelling, “Towards a better baseline for NLP methods in domain-specific information retrieval,” in *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop,* Wien, Österreich: Centromedia, 2005.

[6] E. Voorhees, “Query expansion using lexical-semantic relations,” in *Proc. of SIGIR,* Dublin, 1994.

[7] P. Buitelaar, D. Steffen, M. Volk, D. Widdows, B. Sacaleanu, Š. Vintar, S. Peters, and H. Uszkoreit, “Evaluation Resources for Concept-based Cross-Lingual Information Retrieval in the Medical Domain,” in *Proc. of LREC,* Lissabon, 2004.

[8] L. Ballesteros and B. Croft, “Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval,” in *Proc. of SIGIR,* 1997, pp. 84-91.

[9] L. Ballesteros and B. Croft, “Resolving Ambiguity for Cross-Language Retrieval,” in *Proc. of SIGIR,* 1998, pp. 64-71.

[10] T. Hedlund, E. Airio, H. Keskustalo, R. Lehtokangas, A. Pirkola, and J. Kalervo, “ARVELIN: Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000–2002,” *Information Retrieval,* 7, 99–119, 2004.

[11] M. Braschler, “Combination Approaches for Multilingual Text Retrieval Eurospider Information Technology,” *Information Retrieval,* 7, 183–204, 2004.

[12] N. F. Noy and D. L. McGuiness, “Ontology Development 101: A Guide to Creating Your First Ontology,” SMI Technical Report SMI-2001-0880, 2001.

[13] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne, “Okapi at TREC-4,” in *NIST Special Publication 500-236, the Fourth Text Retrieval Conference (TREC-4)*, 1995, pp 73-96.

[14] TREC http://trec.nist.gov/ 26-12-2006