

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).

