A Cross-language Information Retrieval Based on an Arabic Ontology in the Legal Domain

S. Zaidi
University of Annaba
zaidisor@gmail.com

M.T. Laskri
University of Annaba
laskri@univ-annaba.org

K. Bechkoum
University of Wolverhampton
k.bechkoum@wlv.ac.uk

Abstract

In this paper, we describe a web-based multilingual tool for Arabic information retrieval based on ontology in the legal domain. We illustrate the manual construction of the ontology and the way it is edited using Protégé2000. Using Arabic (UN) documents we identify the legal terms and the semantic relations between them before mapping them onto their position in the ontology. The process of search for Arabic documents uses a query expansion. In addition to finding related documents in Arabic, the retrieval process is further enriched by enabling the user to translate the query into another language, The query expansion is achieved using a semantic word thesaurus namely, Wordnet, which is available on the Web. with a view to retrieve documents in these languages too. A set of query words is used to enable the machine translation of the query from Arabic into English and from Arabic into French.

Keywords : Cross-language information retrieval, Arabic ontology, légal domain, Protégé2000, query expansion, Wordnet.

1. Introduction

The task of finding relevant information amongst large, multilingual and domain-specific collections of text is an active field of research[1]. An impressive and overwhelming quantity of information is readily available on the Web. The challenging question becomes not *where to find* the information, but rather *how to find it.* It is very much *like looking for a needle in a haystack.*

The search for information is often performed using automatic tools such as search engines, directories or meta-engines. These tools are based, in their search, on key words the semantics of which are omitted, thus generating a great number of irrelevant documents.

In the approach presented here, an attempt is made to improve the precision of the search thus minimising the level of noise in the results. Precision here is expressed by the number of relevant documents divided by the total number of documents found. A process of a query-expansion, using an Arabic ontology in the legal domain, achieves this improvement of precision. The ontology is built manually using the ontology editor: Protégé2000. In addition to finding Arabic documents, the information retrieval process is enriched by enabling the user to retrieve English or French documents too. In order to do so the original query is translated (using machine translation) into the target language, French or English. The translated query is then extended using Wordnet, a lexical base that is freely available on the web.

In section 2 we give a general overview on web information retrieval (IR). Section 3 presents the key methods and techniques and the related issues while section 4 focuses on Arabic IR. The general architecture of our system is described in section 5. A walk-through example and concluding notes are to be found in Section 6.

2. Information retrieval on the web

The volume of documents that is created, stored, or needing to be managed on the web, does not cease growing. The number of pages created daily on the web is estimated to be in the region of a million or so. Finding the right information amongst this massive data overload is a challenging task. A task that a mere use of search engines cannot resolve. One might even question whether search engines contribute to the solution or are part of the problem of making Information Retrieval

just that little bit more complicated.

Indeed, the majority of search engines such as Google, Alta Vista, Hahooa and Ajeeb are based on key words, with little account of the semantic of these words.

The user is overwhelmed by an overload of irrelevant information, referred to as *noise*. Sometimes the system (in this case the search engine) does not return relevant documents present in the collection. This is often called *silence*. Because of the great increase in the volume of electronic information, the challenge lies in creating tools to facilitate access to the relevant documents wherefrom desired information is extracted quickly and accurately[2].

3. Information retrieval: key methods

3.1 Documents categorisation

The categorisation of documents consists of assigning documents to pre-set categories according to their contents. The goal of document categorisation is to enable all similar documents to be identified based on a particular query[3]. An automatic categoriser was developed by Sakhr for Arabic[7].

This method may contribute to alleviating the problem of *silence* but that of the issue of *noise* remains always posed.

3.2 Query expansion

The basic principle here is to extend the query by adding new words deemed to be somehow (usually semantically) connected to those contained in the initial query. If there is a strong relation between *word1* and *word2*, whilst *word1* appears in the initial query, then we can replace *word1* by (*word1* $\vee$ *word2*). *word2* is considered as a synonym of *word1*. To extend the query, we generally use a dictionary of synonyms, a thesaurus, or an ontology. The query expansion has an advantage to avoid as much as possible *silence* and improve precision. The main disadvantage lies in the cost of the manual construction of an ontology, and its representativeness of the specified domain.

3.3 The read path method

In this method the granularity is not the page but the zone of text wich consists of one or several paragraphs. The response to a query will be a virtual document made of a series of zones of texts considered to be relevant. The advantage here lies in taking the zone of text like granularity, making it possible to get rid of the nonrelevant zones, which are on the same web page. The disadvantage, according to Radouani [6], is that if the zone of text contains only one paragraph (very small vector) the measurement of their similarity is not easily usable and the results are poor.

3.4 The automatic summary

Some search tools generate a summary of the documents that are good candidates for being relevant. The user can then discard some of the obvious 'noises' thus alleviating the need to download irrelevant documents from remote locations. The summary is given with the key words highlighted in a different colour. This method is very useful in enabling the user to decide about the relevance of information recalled without having to read the entire document. The main drawback stems from the inherent nature of Natural Language Processing (NLP) itself namely, document' segmentation, analysis, and generating summaries.

3.5 Question- Answer query

The query is a simple question in natural language, the response by the system is as precise as the question. This system is composed of three components:

- Analyser of question (generation of a list of words).
- Extraction of the named entities (extracted the turned over documents).
- Extractor of answer (determination of the relevant answers starting from the named entities using the question-type and a list of key words).

The drawback of this method is that the question must have a standard format, which can be very restrictive to novice users.

4. Arabic IR

Arabic is one of the six UN official languages. It is a Semitic language and a mother tongue of up to 150 million people in 21 Arab countries. The number of Arab net surfers in 2002 was estimated to be about 4.4 million. The Arabic alphabet consists of 28

characters. Three of these characters appear in different shapes as follows:

- Hamza (ء) is sometimes written : أ , إ or ا (alif)
- Ta marbouta (ة) like t in english found at the end without two dots (ه = ha)
- Alif maqsurah (ى) is the character (ي = ya) without dots.t

The above three characters pose some difficulties in the setting up of IR system. Some information centres ignore the *hamza* and the dots. above *ta marbouta* to unite the input and output for these characters. There is in Arabic a whole series of non-alphabetic signs, added above or below the consonant letters to make the reading of the word less ambiguous. These are called vowels, or diacritical marks, and are of four types:

- فتحة (ـَ) above a consonant, pronounced like a in cat
- ضمة (ـُ) above a consonant, pronounced like u in put
- كسرة (ـِ) below a consonant, pronounced like the i in big
- سكون (ـْ) above a consonant, pronounced like o in factory

There has been a number of research projects that attempted to design Arabic information retrieval systems. One of these is the MicroAirs system by Al Kharashi in 1991[15]. Al Kharashi's work was an investigation of the three main search methods namely, word stem or root with a view to identify the most suitable for Arabic. His study reveals a superiority of root retrieval methods over the stem or world retrieval method in performing recall but the precision is better with word method. A similar study was done by Abu Salem in 1992. Abu Salem repeated Al Kharashi's experiments comparing the use of words, stems and roots and found that the stem and root retrieval methods perform better than the word one. At lower recall levels the root retrieval method does not perform better than the stem, however at high recall levels the root performs better than the stem[10].

AIRSMA (Arabic Information Retrieval System based on Morphological Analysis) is an another study carried out by Al Tayyar [1]], [11]for his PhD thesis. Al Tayyar compares four methods: Root, Stem, Word and the morpho-semantic method, which he developed. The results of the study seem to suggest that both the root and the stem retrieval methods give better performance than word and morpho-semantic methods in terms of precision. On the other hand, in terms of recall, the root and the morpho-semantic methods perform better[11].

T. Rachedi et al. Developped *Barq system,* which is a distributed multilingual search engine designed specifically for the Arabic language. In order to avoid the drawbacks of the *word* retrieval method, Rachedi et al. used a query expansion based on three concept thesauri. The query is extended to contain the root and the concepts extracted from the thesauri for each non-stop word query term. The results show that the system improves only the recall[12].

5. The proposed cross-language system

5.1 Aims and objectives

Our aim is to improve both *recall* and *precision* of Arabic information retrieval in the legal domain. For this we used an Arabic ontology in the legal domain. The originality of the method presented here stems from the combination of both the root and the word retrieval methods in a way that enables an improvement of both the precision and the recall.

In order to work towards this aim we propose a solution that is a query expansion, based on an Arabic ontology in the legal domain (mainly the Algerian legislation, which draws its laws, in particular the civil code, from the French civil code) [13]. The cross-language retrieval proves to be considerably important in enabling the sharing and distribution of information, independently of the language used or indeed the format in which the information is presented. With this in mind, we developed this multi-linguistic tool, in order to allow the user to access documents in the language of their choice (Arabic, French or English).

5.2 Ontologies

One of the widely used definitions of an ontology is that given by Gruber [4]. *An ontology is a specification of a conceptualisation.* A conceptualisation is the whole of the existing entities in the field and the existing relations between these entities. An ontology defines the vocabulary and the

associated semantics which make it possible for two agents to communicate on a given field of knowledge.

For a better understanding of Gruber's definition, Guarino clarifies the concept of conceptualisation as being "the identification by terms and/or symbols, concepts of the field and existing relations between these concepts" [5].

5.3 Why the legal domain?

It is one of the three pillars of the Algerian constitution. Furthermore a large community (of potential users) works in the legal field namely, judges, prosecutors, lawyers, notaries, bailiffs, academic researchers, and students. Existing Arabic information retrieval systems are generally still legging behind but this is more so for the case of tools specifically designed for the legal domain. A contribution towards a remedial course seems to be quite justified. This is particularly crucial as access to legal information is a requirement on every individual.

5.4 The ontology construction

We have attempted to construct an Arabic ontology in the legal domain according to steps proposed by Noy and Mcguiness [13]. We have used a top-down strategy for constructing the hierarchy of concepts starting by (النظام القضائي_ = legal system) and (القانون_المدني= civil law). The relationship between these two concepts is an ***is-a***. A class A ***is-a*** class B if all instances of B are also instances of A. For example (محكمة عنابة= court of Annaba) is an instance of (محكمة_ولائية _= departmental court) then it is also an instance of a class (محكمة= court) because (محكمة_ولائية= departmental court) is a subclass of (محكمة= court). To each concept we associate its synonyms together with a restriction of derivatives set, which are the most strongly related to the legal domain. We thus combine the benefits of both approches, root and word. For populating our ontology we have used UN articles in Arabic language [8] and articles from a selection of Arabic newspapers. When a query is activated we match the query terms with the concepts of ontology as well as the synonyms and associated derivatives. These derivatives are most strongly related to the legal domain and are determined by a statistical analysis. The latter analysis searches for words often used in the legal domain, using an dedicated legal Arabic corpus. By doing so, *precision* is improved using the word method while the derivatives enable an improvement of *recall*.

Each concept has slots or attributes that determine its properties. If a slot has constraints we define its facets (such as type, cardinality, allowed values etc). We use Protege-2000 for editing our ontology because it is the most powerfull tool, which supports the Arabic language.

5.5 The general architecture of proposed system

The system that we propose to improve the arabic information retrieval on the Web in the legal domain, is situated in a general architecture of an Arabic search engine supporting the translation in English or French queries. The aim is to return documents written in Arabic, French or English.

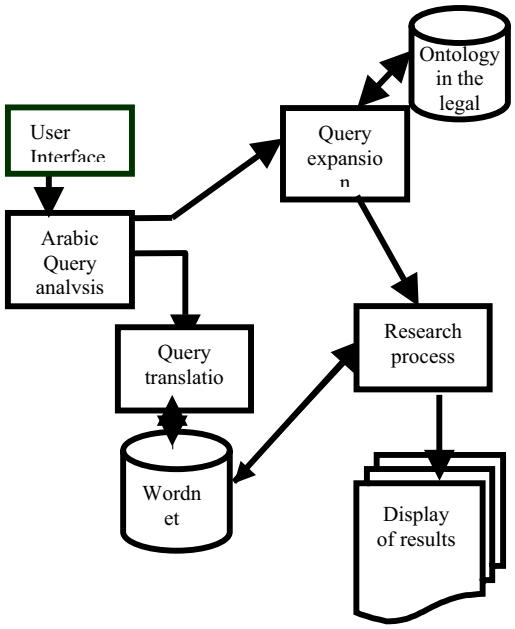

Figure 1. The general architecture of the proposed system

The *user interface* is simple Direct Manipulation Metaphor is used here. A query is formulated using a set of keywords and introduced by the user.

Within the *query analysis* part the query words are separated. '*S*topwords' such as: (في = in), (من= from), (كـ= as), (لـ=for), (مثل = like) are eliminated. Example of a query حقو ق الطفل في الوطن العربي) = child rights in Arab

countries). We keep (حقوق الطفل الوطن العربي = rights, child, countries, Arab). The reason for eliminating the 'stopwords' is that these words are not indexed and they do not add any crucial information. After that the query undergoes normalization as follows: the punctuation, all which is nonletter, diacritics marks are removed, (أ, إ , آ) are remplaced by (ا), (ة) is remplaced by (ه) and the final (ى) is remplaced by (ي).

Within the process of *Query expansion* we navigate through ontology, looking for each term of the query, we add synonyms, relative derivatives and generic or specific concepts to the initial query for extending it. This is done with a view to avoid a possible silence, and to perform the precision. In the above example the query becomes: (حق ،الحق، حقوق، الحقوق = a right, the right, rights) and الاطفال،اطفال،طفوله، (الطفل، طفل الطفوله = a child, the child, children, childhood) and (دوله،دول،الدوله وطن، الدول،بلد،البلد،بلدان،البلدان الوطن،الاوطان، = a country, the country, countries, a nation, the nation, nations, a state, the state, states) and (الاسلاميه، اسلاميه عربي، العربي،اسلامي،الاسلامي =, an Arabian, the Arabian, an islamic, the islamic). In the *search process* the new extended query is submitted to a search tool (or any search engine that supports Arabic query) which returns a list of documents classified according to their relevance algorithm. The document is considered relevant if it has the greatest number of occurrence of the query' words. The boolean operator "and" is included between diffrent sets of concepts.

Query translation is used if the user wishes to obtain French or English documents. In this case they will choose the machine translation of the query in the target language, then their query is extended with Wordnet before the search is carried out.

6. Discussion and Concluding Remarks

In what follows we make a brief attempt to show the difference between a simple query and our extended query, using the ontology of the legal domain.

Initial query: (قانون العقوبات = penal code) Results of research with the Arabic engine "Hahooa " [14]:

Recall	Precision
115	2 out of the first ten documents

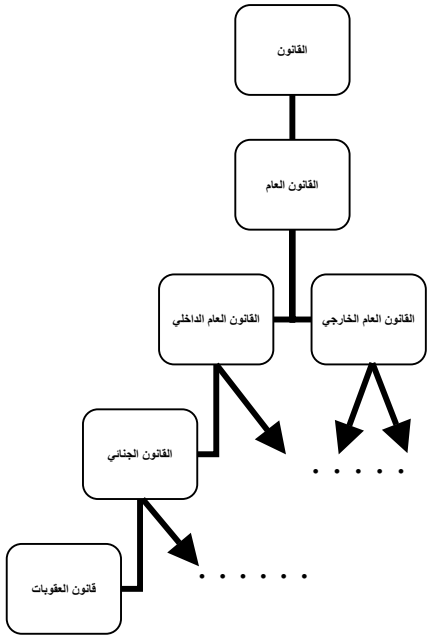

Figure 2. The hierarchy of the concepts

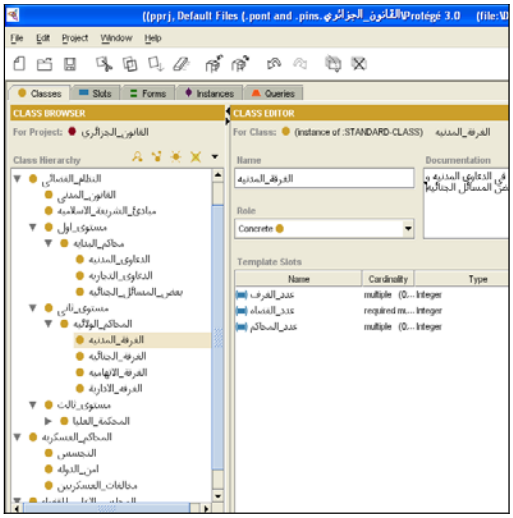

Figure 3. Presentation of the concepts by using protégé-2000

- Query expansion

Extended query: we search in the ontology the concepts**: (قانون = code and العقوبات =penal)** then we take all the synonyms and the relative derivatives attached to the concept. when it is necessary we take also hyponyms or hyperonyms. The extended query becomes:

قانون العقوبات، قوانين ،عقوبه ، العقوبه، يعاقب، " نص،نصوص، تشريع، حكم ،احكام، مخالفة، مخالفات، جنحة، جنح، تخريب، اعتداء،اعتداءات، قتل ، "جريمة،جرائم، قانون _ عام _ داخلي

The first words are synonyms of the query words; the last is a generic concept. Results of search with the Arabic engine "Hahooa ":

Recall	Precision
135	7out of the first ten documents

- *Translation of the query in English*

Using Tarjim of Ajeeb (available on-line by Sakhr company software) [7], we use machine translation and we obtain ***penal code***

- *Query expansion with wordnet*

After entering penal code in Wordnet we obtain:

The noun "penal code" has 1 sense in WordNet.
1. Penal code -- (the legal code governing crimes and their punishment)

- *Search processing*

Then we use the extended query and submit it to search engine. The outcome is summarized below:

Recall	Precision
1230	7out of the first ten documents

The preliminary results are quite promising and show that there is a significant improvement in the recall, and the precision.

We presented in this paper a method of a query expansion based on an ontology in the legal domain, in Arabic language. We intend further this work by calculating the relevance on the first fifty documents and to building our own tool of search. The relevance of the algorithm used by this tool will be based on the frequency of the initial key words which will have priority compared to words released by ontology.

References and bibliography

.

[1] M. Al-tayyar, & Bechkoum, K. (1998) "The Effectiveness of Morphological Analysis for Text Retrieval in Arabic", 6th International Conference on Multi-lingual Computing, Cambridge, UK, 17-18 April 1998.

[2] C. Boudry, « Typologie et mode de fonctionnement des outils de recherche d'information sur internet en biologie/médecine » parue dans MEDECINE/SCIENCES 2002.

[3] M. El Kourdi, A. Bensaid, T. Rachidi "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm" 2004.

[4] T. Gruber "what is an ontology?" (http://ksl-web.stanford.edu/people/gruber/)

[5] N. Guarino "Understanding, Building, And Using Ontologies" *Oct 1996* LADSEB-CNR, National research Council Corso Stati Uniti 4, I-35127 Padova, Italy

[6] S. Radhouani, J.P. Chevallet, M.Géry. « Extraction et indexation de chemins de lecture pour la recherche d'information sur le web » 2002.

[7] Sakhr software company: (www.sakhrsoft.com) 2004.

[8] united nation development programme (www.undp.org)

[9] B. Wang, H. Xu, Z. Yang, Y. Liu, X. Cheng, D. Bu, S. Bai "*Filtering, Web and QA"* TREC-10 Experiments at CAS-ICT Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China 2001.

[10] H. Abu Salem "A microcomputer based Arabic bibliographic information retrieval system with relation thesaurus (Arabic-IRS), Ph.D Thesis Chicago Illinois Institute of technology.

[11] M. S. Al Tayyar "Arabic information retrieval system based on morphological analysis (AIRSMA)" Ph. D. Thesis DeMonfort University july 2000.

[12] T. Rachedi, and al., "Barq: distributed multilingual Internet search engine with focus on Arabic language," in proc of IEEE conf. On Sys., Man and Cyber., Washington DC, Oct. 5-8, 2003.

[13] N. F. Noy and D. McGuinness «Ontology Development 101: A Guide to Creating Your First Ontology»

[14] Hahooa, http://www.hahooa.com/nav.php?ver=ar

[15] I. Al Kharashi A microcomputer-based Arabic information retrieval system comparing words, stems, and root as index terms, Ph.D.Thesis.Chicago:Illinois Institute of technology.