IMPROVING Q/A USING ARABIC WORDNET

Lahsen Abouenour[1], Karim Bouzoubaa[1], Paolo Rosso[2]

[1] Mohammadia School of Engineers - Med V University - Agdal, Rabat – Morocco
abouenour@yahoo.fr, karim.bouzoubaa@emi.ac.ma

[2] Natural Language Engineering Lab., Universidad Politécnica Valencia, Spain
prosso@dsic.upv.es

ABSTRACT

With the expansion of the content available on the web, Question/Answering (Q/A) systems have become, among other searching tools, a focus of researchers and users as well. For the Arabic language very few works have been done in this field. In this paper, we focus on the improvement of Q/A through a Query Expansion (QE) process. Our approach is based on the ontology that we have built using Arabic WordNet. Indeed, we designed a QE process from the semantic relations existing among the concepts of our ontology. The preliminary experiments that we have conducted show that the accuracy of getting the answer expected was improved by our QE approach.

***Keywords**: Question/Answering, Query Expansion, Ontology, Arabic WordNet, Semantics, Morphology*

1. INTRODUCTION

The great amount of information available on the web has become an important resource for knowledge-based economies. However, in this context, users tend to be lost when seeking for specific information. Tools such as Search Engines (SEs), Information Retrieval (IR) systems and Question/Answering (Q/A) systems [14] have been developed and are being essential to help users in their searching processes.

SE, IR and Q/A systems are three different kinds of Natural Language Processing (NLP) applications. Researchers in these fields are faced to both the complexity due to the kind of NLP applications and to that of the language in which information are written. The need of such systems is higher in the context of the Arabic language which presents many challenges to the research community regarding its particularities (short vowels, absence of capital letters, complex morphology, etc.). When we go deeply in the details, the three kinds of searching tools are different. Indeed, unlike the traditional IR systems, SEs operate in an open environment (the Web) and their users are not experts. Therefore, additional techniques are added to make SEs like interactive IR and try to get relevant results from a set of unstructured content. The case of Q/A is particular in that the question is in natural language instead of a set of keywords. Moreover, we expect an answer (or combination of answers) instead of a set of documents without requiring the user's manual and sometimes much tedious filtration. Consequently, Q/A systems make the exploitation of the results much easier than SEs and IR systems. However, the aim of the three tools is common: searching relevant information in a set of documents. Therefore, at a higher level their processing is the same: the user enters a query, and then the system extracts keywords from the query, expands them with other related keywords, search in the target documents and returns the result to the user. It is of evidence that one of the most important modules that the three kinds of searching tools have is Query Expansion (QE). Indeed, it allows the search of documents that the system based of original keywords would not consider. Classically, this expansion is done at the morphological level. For example, for the keyword معرفة (mErfp[1], knowledge) that would appear in a user query the QE module extends and provides other morphological forms that the system can use such as : عارف (EArf, who knows), معارف (mEArf, plural of knowledge), معرف (mErf, get known), ...

In this paper, we focus on the QE process in the context of Arabic NLP. Our attention will be devoted to its use in the context of Q/A systems which is, for the Arabic language, still less concerned by NLP researchers. Our approach is to expand the query of the user not only at the morphological level but at the semantic one also. For this purpose, we make use of the Arabic WordNet (AWN).

The structure of the article is as follows. In the second section we present some related works to Arabic Q/A systems and QE techniques. Then, in the Section Three we describe the steps we followed to build our linguistic resource which is Amine AWN (AAWN). After that, we present through an example the usefulness of our semantic approach in the improvement of Arabic Q/A and we give some results related to preliminary experiments on a set of CLEF[2] questions. Finally, in the last section we draw some conclusions and the future works to be done.

[1] We use the buckwalter transliteration available at http://www.qamus.org/transliteration.htm

2 Cross Language Evaluation Forum, http://www.clef-campaign.org

2. RELATED WORKS

Unlike Information Retrieval applications, Question/Answering systems try to obtain a simple answer to a specific question (with both the question and the answer being formulated in natural language). The study of the Q/A task research guidelines [4] reported that there are generally four kinds of questioners where each type represents questions with a certain level of complexity:

(i) Casual questioner: asking concrete questions about specific facts;
(ii) Template questioner: this type of questioner might ask some questions which require the system to retrieve portions of the answer from different documents and combine them in just one answer;
(iii) Cub reporter: this other type of questioners would require a Q/A system able to collect many information from different sources about a single fact;
(iv) Professional information analyst: finally, this is the highest level of questioners which need a system able to deduce and decide by itself the answer.

The basic Q/A processing cycle is composed of three major stages [7]:

(i) Processing the input question
(ii) Retrieving with an IR system the candidate documents (paragraphs) containing the answer
(iii) Processing each one of the candidate documents (paragraphs) in the same way as the question is processed and returning those sentences that may contain the answer.

The first stage needs a question classifier, a query expansion module (for keywords) and a named entity recognizer. The latter is required also in the third stage. Just few implementations of Arabic Q/A systems exist:

- QARAB [7] is a system that takes natural language questions expressed in the Arabic language and attempts to provide short answers. The system's primary source of knowledge is a collection of Arabic newspaper text extracted from Al-Raya, a newspaper published in Qatar. QARAB uses shallow language understanding to process questions and it does not attempt to understand the content of the question at a deep, semantic level.
- AQAS [10] is knowledge-based and, therefore, extracts answers only from structured data and not from raw text (non structured text written in natural language).
- ArabiQA [2] is an Arabic Q/A prototype based on the JIRS [3] Passage Retrieval (PR) system and a Named Entities Recognition (NER) module. It embeds an Answer Extraction module dedicated especially to factoid questions. In order to implement this module authors developed an Arabic NER system [1] and a set of patterns manually built for each type of question.

Generally, for Arabic, the existing Q/A systems fail when a complex question has to be processed. A complex question is characterized by the need of domain knowledge, and there is no single answer type that can be identified, but rather an answer structure needs to be recognized.

For instance the question: إلى أي مدى ساهم ارتفاع أسعار النفط في الرفع من تكلفة المعيشة ؟ (To what extent the increase of oil prices has raised the cost of living?) refers to the economy domain. We need question decomposition in order to get the structure of the answer:

- What is meant by تكلفة المعيشة (the cost of living)?
- How can the system link the expression ساهم في الرفع (contribute in increasing) to the concept "impact"?
- How does one define the increase or decrease of a problem?

Indeed, complex questions (i.e., of a professional information analyst level of complexity) need to be decomposed into a set of simpler questions by the adoption of a knowledge base question analyzer and answer extraction module. Thus, a semantic query expansion can be useful in the three stages of a Q/A system. In this paper our approach is partially similar to the one for English described in [16] where authors have built a module for extracting the final answer from retrieved documents in a Q/A system using WordNet. Their approach is related to extended unification based on ontology (WordNet) in the third stage of the Q/A processing. In our work we start the semantic expansion in the first stage and the generated keywords can be used in the other stages. In addition, our approach differs from the one of [16] because: (i) it uses the concepts of SUMO and their definitions; (ii) it uses the content of WordNet within a Platform of Artificial Intelligence.

In the next section, we describe the steps that we followed to build the Amine Arabic WordNet ontology which is the kernel of our work.

3. BUILDING THE AMINE AWN ONTOLOGY

Amine[1] is a Java open source multi-layer platform dedicated to the development of intelligent systems and multi-agents systems [8]. It is a modular environment composed of four layers: (i) Ontology layer; (ii) Algebraic layer; (iii) Programming layer; (iv) Agents and Multi-Agents Systems layer. We recommend the reader to consult the web site[3] of Amine for further details.

[3] http://amine-platform.sourceforge.net

In Amine the definition provided by John Sowa[4] of an ontology as a “catalogue of types” (which is organized in a hierarchy, called type hierarchy and can be considered as a first and basic class of ontology *type hierarchy ontology*) is extended. Indeed, to be able to use more general classes of ontology, the assumption in Amine is the possibility to associate to each type (category) all (or most of) the knowledge acquired by the system concerning this type. Such knowledge is organized in terms of Conceptual Structures (CSs): type definition, canon for a type and schemata (called situations in Amine). Individuals (instances) are associated to their types [9]. Figure 1 shows the different ontology classes supported by Amine.

Building an Arabic ontology is not a simple task. For doing so, we need semantic resources. In comparison with other languages, not many are the NLP tools and resources in general (corpora, gazetteers, etc) which are available for Arabic [13]. This is especially true for semantic resources. Recently, this picture is about to change with the new release of Arabic WordNet.

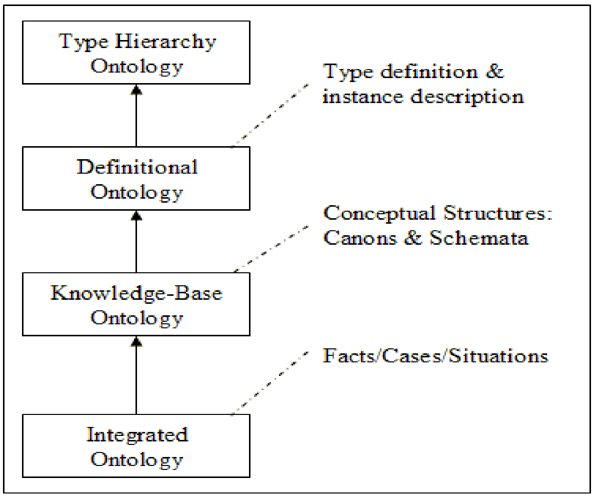


Figure 1: Ontology classes

AWN is a free lexical resource for modern standard Arabic [5]. It is based on the design and contents of Princeton WordNet (PWN) [6] and can be mapped onto PWN as well as a number of other wordnets, enabling translation on the lexical level to and from dozens of other languages. Moreover, the mapping of WordNet to the Suggested Upper Merged Ontology (SUMO) [11] [12] provides opportunities to use the semantic side in some Arabic NLP applications.

Thus, the idea is to combine the richness and the accuracy of the Arabic WordNet with the characteristics of the Amine Platform by building the Amine AWN ontology (Figure 2).

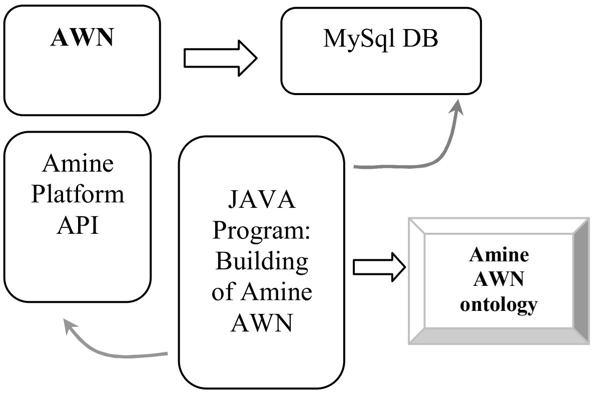


Figure 2 : Building of Amine AWN ontology

The procedure that we followed to accomplish this task can be glossed as follows: we exported the entire data embedded in AWN to be used by a Java module based on Amine Platform APIs. This program uses the mappings existing between English synsets (WordNet) and SUMO concepts to build the Amine AWN *type hierarchy*. Then, it adds Arabic synonyms to those types based on the equivalence relation between English synsets and Arabic WN synsets.

The equivalence is not the only relation which links WordNet synsets to the SUMO concepts. Indeed, in WordNet a synset can be more specialized in which case our module creates a new *subtype* of the SUMO concept. At that moment, Arabic synsets are added as synonyms for the new entry (*subtype*) created. In case of *has instance* a new *individual* is created instead of a *subtype*. At this moment we have a first level of our *type hierarchy*.

The second level is obtained by a similar processing where relation between English synsets is exploited. Therefore, at this stage a *hyponymy (or hypernymy)* relation takes the role of a *specialization (or generalization)* relation of the previous stage. In addition, our module allows the automatic extraction of SUMO concepts definitions written in SUO-KIF notation: therefore, the conversion to the CGIF notation (supported by Amine Platform) is to be done. For the purpose of the present article we manually added definitions for the concept types concerned by the example of the next section.

4. EXPLOITATION OF AMINE AWN IN A Q/A SYSTEM

As we mentioned in the section above, a simple processing of a question (based on question type patterns and NER) is not enough to deal with complex questions.

From *the Worldwide Islamic Network of Women* web site[5] we take the following passage:

[4] http://www.jfsowa.com/ontology/index.htm

[5] http://www.islamwomen.net/ArbIW/ACRMWDetails.aspx?id=48

خلفية مهنية : عضو مجلس الأعضاء لمؤسسة التعاون بجنيف/عضو مؤسس للمؤسسة الثقافية العربية بلندن./عضو جمعية الصحافيين و الخريجين الكويتية، ورابطة الأدباء، وجمعية الاقتصاديين الكويتية. عضو اللجنة التنفيذية لمنظمة حقوق الإنسان في الوطن العربي (عضو مؤسس)./عضو المجلس الاستشاري الأعلى للتربية – الكويت. /عضو اللجنة التنفيذية للمنظمة العالمية للنساء المسلمات لجنوب شرق آسيا/عضو الاتحاد العالمي لاقتصاديات الطاقة/عضو مجلس الأمناء واللجنة التنفيذية لمنتدى الفكر العربي في عمان. /عضو مجلس الأمناء بمركز الدراسات العبرية- جامعة اليرموك. /عضو مساند بمركز الدراسات العربية ببيروت/عضو مجلس إدارة مشروع بحوث الشرق الأوسط والمعلومات بواشنطن./عضو مؤسس للمجلس العربي للطفولة والتنمية بالقاهرة. /عضو جمعية علم الاجتماع العربية بتونس. /عضو المجلس الاستشاري للاتحاد الدولي لتنظيم الأسرة بلندن / .

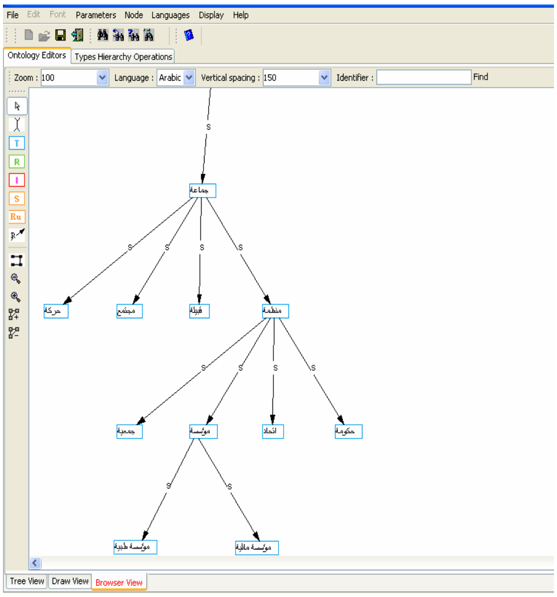
This passage is a part of Souad Assabah's biography and represents the content from which we process the following question: ماهي المنظمات التي كانت سعاد الصباح تشتغل فيها ؟ (i.e., What are the organizations that Souad Assabah was working for ?). The keywords of our question are: سعاد الصباح - تشتغل - المنظمات (organizations, work, Souad Assabah). With a classical Q/A system based on a morphological query expansion we can add some other forms of the keyword المنظمات such as : منظمة (mnZmp, organization), تنظيم (tnZym, to organize), نظام (nZAm, system) . Unfortunately, these keywords are not enough to extract the whole answer from the passage. Therefore, the system, as it is, will return a small part of the expected answer and will fail to get the other parts of the answer from the passage.

Let us see if using Amine AWN ontology could help to get the other parts of the answer. According to the structure of Amine AWN ontology we can move from one concept to another using the following semantic Amine AWN links: (i) Concept synonyms; (ii) Concept supertype; (iii) Conceptual structure definition; (iv) Concept subtypes.

The first semantic expansion (by synonyms) processing gets the synonyms of the word منظمة from Amine AWN ontology. Indeed, there is the keyword: تنظيم. This keyword does not occur in the passage. Therefore, we move to the next step of our processing. Figure 3 illustrates the position of the Organization concept in Amine AWN hierarchy.

The second semantic expansion (by supertypes) shows that the Organization concept has a more general type (*supertype*) which is جماعة (jmAEp, community)**.** Also in this case the passage above does not contain this new keyword.

The third semantic expansion (by definition) that we process is based on the definition of the concept (Organization) in Amine AWN ontology which is: "An &%Organization is a corporate or similar institution. The &%Members of an &%Organization typically have a common purpose or function". Note that the symbol &% shows that the concept exists in our ontology. Thus, the concept Organization has a semantic link with an other concept of the Amine AWN ontology which is Member (EDw ،عضو). Now we reach a new keyword which appears in the passage above.



Zoom : 100 Language : Arabic Vertical spacing : 150 Identifier : Find

Figure 3: Position of the Organization concept within Amine AWN hierarchy

The final semantic expansion (by subtypes) is related to the subtypes of the Organization concept. In AWN we find new keywords like : حُكومَة - اِتِّحَاد - وَفد - مُؤَسَّسَة - جَمْعِيَّة – وَحْدَة

Among these new keywords three appear in our passage: اِتِّحَاد-مُؤَسَّسَة – جَمْعِيَّة

Let us see what happens when applying the same process recursively to each reached concept. Figure 4 illustrates the results of such approach. In this figure boxes with labels 1, 2, 3 and 4 refer to the first, second, third and fourth semantic expansion.

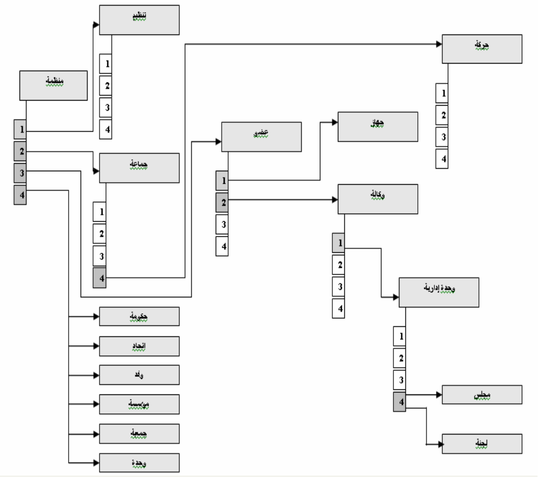


Figure 4: Semantic query expansion processing based on Amine AWN ontology

As we can see, we could expand the query (keywords) semantically and we success to reach words that exist in the given passage: مَجْلِس (mjls, council) and لَجْنَة (ljnp, commission)**.** The keywords provided by our semantic query expansion are useful for the different stages of a Q/A processing. Therefore, we propose the following expansion model to be integrated with a Q/A system (see Figure 5).

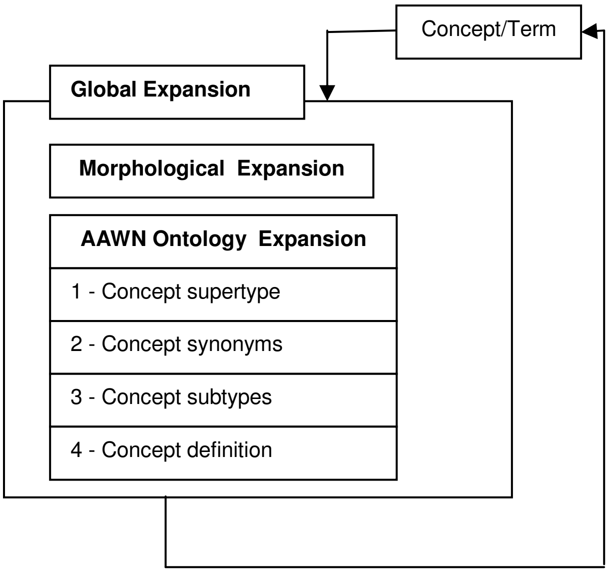


Figure 5: Expansion model to be integrated with a Q/A system

In the next section we present some results of the preliminary experiments using our approach.

5. PRELIMINARY EXPERIMENTS

This section describes preliminary experiments that were undertaken to confirm the effectiveness of our semantic QE in the context of Q/A systems.

As preliminary experiment we have taken a set of 82 questions of the CLEF that was translated into Arabic[6]. These CLEF questions are classified into different domains (sport, geography, politic, etc.) and different types (questions seeking for time answers, persons, places, etc.).

After producing a set of new Keywords based on Amine AWN, we look for the correct answer in the first five snippets returned by Google[7].

Answer	Question
بشار الأسد	من هو رئيس سوريا؟
By Synonym	شيخ - عريف - زعيم - مَكَانَة - مَرْتَبَة - رُتْبَة
By Definition	منصب - وظيفة - -
By subtypes	عام - مدير - أميت عام
By supertypes	قائد - رئيس الدولة - مَرْكَز
مالكوم اكس	من أسس منظمة"مؤسسة المسجد الاسلامي"؟
By Synonym	تنظيم - جِهَاز
By Definition	Organization - عضو -
By subtypes	حُكُومَة - اِتِّحَاد - وَفد - مُؤَسَّسَة - جَمْعِيَّة - وَحْدَة - حَرَكَة
By supertypes	جماعة - وَكَالَة - مَجْلِس - لَجْنَة
الارجنتين	ما هو الفريق الذي واجه فريق البرازيل في أول مبارة عالمية له؟
By Synonym	لاعب - قابل - لاقي - مُشَارِك - مُنَافِس
By Definition	- SocialInteraction - CognitiveAgents - Agent Patient
By subtypes	تَحَدَّى
By supertypes	فعل - مُتَسَابِق
66,60%	ما هي نسبة النمساويين الذين صوتوا لصالح انضمام النمسا الى الاتحاد الاوروبي؟
By Synonym	اِنْتَخَبَ - اِقْتَرَعَ - رَشَّحَ -
By Definition	
By subtypes	تَبَنَّى - اِخْتَار - أَيَّدَ
By supertypes	إِخْتَارَ -

Table 1: Extract of the CLEF questions used in the experiment

In order to evaluate the results, we employed two measures:

- The Accuracy which is the average of the questions where we find the right answer in the first snippet
- The Mean Reciprocal Rank (MRR): The reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer and MRR is the average of the reciprocal ranks of results for a sample of queries[8] [15].

Table 1 lists some examples of the CLEF questions used in this experiment with their QE extension[9] whereas Table 2 illustrates the results obtained:

Domains	# Q	%	Accuracy without AWN	Accuracy using AWN	MRR without AWN	MRR using AWN
History	20	24,69	20,00%	25,00 %	8,55	8,23
Sport	5	6,17	20,00%	20,00 %	10,47	8,00
Politic	12	14,81	25,00%	33,33 %	8,53	10,25
Culture	9	11,11	33,33%	55,56 %	9,44	16,96
Geography	8	9,88	37,50%	50,00 %	11,83	15,67
Technology	7	8,64	28,57%	42,86 %	12,33	14,86
Other	21	25,93	38,10%	23,81 %	11,46	10,14
All	**82**	**-**	**29,26 %**	**32,92 %**	**10,15**	**11,25**
QE by Synonyms	**82**	-	-	**14,63%**	-	**5,42**

[6] http://www.dsic.upv.es/~ybenajiba/downloads.html
[7] www.google.com
[8] http://en.wikipedia.org/wiki/Mean_reciprocal_rank
[9] The set of CLEF questions in Arabic together with their semantic expansion is freely available at: http://www.dsic.upv.es/grupos/nle/downloads.html

QE by Supertypes	**82**	-	-	**12,20%**	-	**5,68**
QE by Subtypes	**82**	-	-	**10,98%**	-	**3,52**
QE by Definitions	**82**	-	-	**7,32%**	-	**2,70**

Table 2: Results of preliminary experiments

As we can see, by using AWN the accuracy has been improved from 29,26% to 32,92%, and the MRR has reached 11,25 against 10,15.

The results show that the stage of QE by synonyms was one of the two most successful semantic expansions with respect to the improvement of both the accuracy and the MRR. Moreover, there are some questions for which the answer does not appear in the first five snippets returned by google. For instance, for the question: ما هي نسبة النمساويين الذين صوتوا لصالح انضمام النمسا الى الاتحاد الاوروبي ؟, the expected answer (66,60%) appears in the first returned snippet only when we extend the keyword صوتوا by subtypes: therefore, when using new keywords such as تَبَنَّى - اِخْتَار – أَيَّدَ.

The results show that for 19,5% of the questions we success to get the expected answer in the first five snippets using our semantic expansion after failing to get it without any QE.

6. CONCLUSION AND FUTURE WORK

As we explained in this paper, the adoption of the Amine AWN ontology traces new ways to get the answer expected by exploiting the definition of a concept, its synonyms and its context represented by the hierarchical nearly types. A more refined model for semantic QE can assign a weight for each produced keyword according to its relation with the source keyword. This weight will be a function depending of the relation type and the distance between the initial concept and the produced keyword.

We have done some preliminary experiments that show the improvement of the possibility of getting the expected answer in the returned documents when using our QE approach. These experiments show also the ability of the system using our approach to get answers where Google fail with respect to the range of snippets considered (the first five). In order to confirm these results, we intend to do other experiments with a large set of questions and using the JAVA Information Retrieval System (JIRS) [3] instead of Google. Measures such as recall, precision and F-measure will be used in the evaluation process.

ACKNOWLEDGEMENT

This research has been done in the context of the following projects: 'AECI-PCI A01031707' and 'CICYT TIN2006-15265-C06'.

REFERENCES

[1] Benajiba Y., Rosso P. "Arabic Named Entity Recognition using Conditional Random Fields", *In: Proc. Workshop on HLT within the Arabic World: Arabic Language & local languages processing Status Updates & Prospects. In the Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, 31st May, 2008.

[2]Benajiba Y., Rosso P., Lyhyaoui A. "Implementation of the ArabiQA Question Answering System's components", In: Proc. Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco, April 3-5, 2007.

[3] Benajiba Y., Rosso P., Gómez J.M. "Adapting JIRS Passage Retrieval System to the Arabic". *In: Proc. 8th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2007*, Springer-Verlag, LNCS(4394), pp. 530-541, 2007.

[4] Carbonell J., Harman D., Hovy E., Maiorano S., Prange J., Sparck-Jones K. "Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization". Rapport technique, NIST, 2000.

[5] Elkateb, S., Black W., Vossen P., Farwell D., Rodríguez H., Pease A., Alkhalifa M. "Arabic WordNet and the Challenges of Arabic". *In proceedings of Arabic NLP/MT Conference*, London, U.K, 2006.

[6] Fellbaum C. 1998. "WordNet: An Electronic Lexical Database". *MIT Press*, *cogsci.princeton.edu/˜wn,* September 7, 2000.

[7] Hammou B., Abu-salem H., Lytinen S., Evens M. "QARAB: A Question answering system to support the ARABic language". *In: Proc. of the workshop on Computational approaches to Semitic languages*, *ACL*, pages 55-65, Philadelphia, 2002.

[8] Kabbaj A., "Development of Intelligent Systems and Multi-Agents Systems with Amine Platform", *In 15th Int. Conf. on Conceptual Structures, ICCS'2006*, Springer-Verlag, 2006.

[9] Kabbaj A., Bouzoubaa K., K. ElHachimi and N. Ourdani. "Ontology in Amine Platform: Structures and Processes", *In the 14th Proc. Int. Conf. Conceptual Structures, ICCS 2006*, Aalborg, Denmark, 2006.

[10] Mohammed F.A., Nasser K., Harb H.M., "A knowledge-based Arabic Question Answering

System (AQAS)". *In: ACM SIGART Bulletin*, pp. 21-33, 1993.

[11] Niles I., Pease A., "Towards a Standard Upper Ontology". *In: Proceedings of FOIS 2001*, Ogunquit, Maine, pp. 2-9, 2001.

[12] Niles I., Pease A., "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology." *In Proceedings of the 2003 International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, 2003.

[13] Rosso P., Benajiba Y., Lyhyaoui A., "Towards an Arabic Question Answering system (in Arabic)". *In: Proc. 4th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROIV*, Damascus, Syria, 11-14 December, 2006.

[14] Roussinov D., Fan W., Robles-Flores J. Beyond Keywords: Automated Question Answering on the Web; Communications of the ACM; Vol. 51 No. 9; 60-65; September, 2008.

[15] Voorhees E.M., "The TREC-8 question answering track report". *In Proceedings of the 8th Text Retrieval Conference*, Gaithersburg, Maryland, USA, pp. 77-82, 1999.

[16] Yarmohammadi Mahsa A., Shamsfard M., Yarmohammadi Mahshid. A., Rouhizadeh M., "Using WordNet in Extracting the Final Answer from Retrieved Documents in a Question Answering System". *In : Proc. The Fourth Global WordNet Conference*, Szeged, Hungary, January 22-25, 2008.