Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system

Lahsen ABOUENOUR
Mohammadia School of Engineers, Med V[th] University
Rabat, Morocco
abouenour@yahoo.fr

Karim Bouzoubaa
Mohammadia School of Engineers, Med V[th] University
Rabat, Morocco
karim.bouzoubaa@emi.ac.ma

Paolo Rosso
Natural Language Engineering Lab, ELiRF, Universidad Politécnica
Valencia, Spain
prosso@dsic.upv.es

Abstract

The adoption of semantic Query Expansion (QE) could be useful in the context of Question/Answering (Q/A) systems. For the Arabic language this is a challenging task since it has many particularities (short vowels, absence of capital letters, complex morphology, etc.). This paper presents an evaluation of a proposed semantic QE based on Arabic WordNet (AWN). Two types of experiments are conducted: the keyword-based evaluation which uses a classical search engine as passage retrieval system, and the structure-based evaluation that uses the Java Information Retrieval System (JIRS) which takes into account the structure of the question. Results show that the best performances in terms of accuracy and Mean Reciprocal Rank are reached when the proposed semantic QE together with JIRS are used.

1 Introduction

With the fast growing of the available content on the web, there is an increasing interest in Question/Answering systems (Q/A) (Carbonell et al., 2000). In fact, classical Information Retrieval (IR) systems are wasteful when users look for a precise answer of a question instead of a set of returned documents.

Unlike other languages such as English, the research in Arabic Natural Language Processing (NLP) has, so far, concentrated less on the Q/A task. Nevertheless, there are a number of attempts to implement automatic Arabic Q/A systems working on structured texts (Mohammed et al., 1993), returning relevant snippets without automatically extracting answers (Hammou et al., 2002), (Benajiba et al., 2007a) as well as recently a semi-automatic Q/A system for factoid questions (Brini et al., to appear in 2009). Most of these systems are based on three modules: question classification and analysis, passage retrieval (PR) and answer extraction. The performance of the latter depends on the results provided by the two first modules. Indeed, if the retrieved passages returned by the second module do not contain the whole or a part of the question keywords, the answer extraction module fails to provide the expected answer.

Most of the time, users concretely formulate the question using words which do not, necessarily, appear in the base documents. Therefore, a Query Expansion (QE) process could be used by the Q/A system modules in order to generate new keywords that may exist in the base documents. Rachidi et al. (2003) cite statistical and dictionary-based QE techniques as the most common for Arabic. These techniques could be useful in the context of Q/A systems. Unfortunately, keywords which are semantically related to the user question may not be provided by a basic QE. Indeed, even if those keywords could be relevant, a QE which uses only lexical and morphological resources might not be able to identify them. Thus, the use of a semantic QE is required since the same question can be formulated using different words with an equivalent meaning. Moreover, the generation of keywords based on the semantic relations makes easier the matching of the question structure and the candidate passages one.

In (Abouenour et al., 2008) we have presented a semantic QE approach with preliminary experiments in the context of the Arabic Q/A task. This approach uses the current release of the Arabic

Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pages 62–68,
Athens, Greece, 31 March, 2009.

WordNet[1] ontology (AWN) (Elkateb et al., 2006; Rodriguez et al., 2008). Let us recall briefly that AWN ontology is a free lexical resource for modern standard Arabic (Elkateb et al., 2006). It is based on the design and contents of Princeton WordNet (PWN) (Fellbaum, 2000) 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. AWN is also connected to SUMO (Supper Upper Merged Ontology) (Niles and Pease, 2001; Niles and Pease, 2003).

Our approach uses not only the current content of AWN but also four of its semantic relations. Indeed, we use a QE process based on: (i) QE by synonyms, (ii) QE by definitions, (iii) QE by subtypes, (iv) QE by supertypes.

In order to be able, in further works, to consider other semantic operations, we have implemented our approach using the Amine Platform[2]. Amine is a Java open source multi-layer platform dedicated to the development of intelligent systems and multi-agents systems (Kabbaj et al., 2006). Thus, the Amine AWN (AAWN) hierarchy is based on the content and the structure of AWN. For each concept type in AAWN there are synonyms, subtypes and supertypes with respect to the synonymy, hyponymy and hypernymy relations in AWN. The implementation of QE by definition uses the SUMO concepts definitions written in SUO-KIF notation.

The added value of this semantic QE in the context of Arabic Q/A systems has been illustrated by examples in (Abouenour et al., 2008). We have conducted preliminary experiments with 82 CLEF[3] questions with manual search using Google Search Engine (SE).

In the community and in order to evaluate the results, two measures are considered:

- 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: MRR is the average of the reciprocal ranks of results for a sample of queries[4] (Voorhees, 1999).

The results have shown that the semantic QE based on AAWN ontology has improved the accuracy by 3,66 % and the MRR by 1,10. Preliminary experiments consider keywords separately without taking into account the question structure. The Java Information Retrieval System (JIRS[5]) (Benajiba et al., 2007b) is a passage retrieval system which can allow us to consider this structure.

Our general aim is to develop a separate semantic QE module that could be used within Q/A systems. In the context of the current paper, our objective is two fold: (i) confirm previous results with large and automatic experiments using the Yahoo API; (ii) take into account the structure of the question and retrieved passages using JIRS.

The structure of this paper is as follows: in Section 2 we give more details about the evaluation and the refinement process of our semantic Query Expansion. Section 3 is devoted to the results of the automatic evaluation based only on keywords. In Section 4, we present the results of the structure-based evaluation. Section 5 is a synthesis of the results reached in the two experiments. Finally, in the last section we draw some conclusions and we discuss the future works to be done.

2 Evaluation and refinement process of our semantic Query Expansion

The tasks related to the design of our semantic Query Expansion are illustrated in Figure 1.

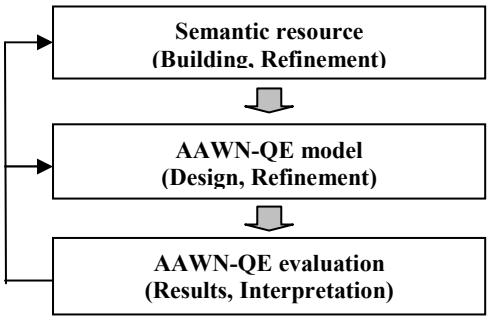

Figure 1. The proposed semantic QE approach

The above figure shows that one of the evaluation process aims is evaluating both the QE

[1] http://www.globalwordnet.org/AWN/
[2] http://amine-platform.sourceforge.net
[3] Cross Language Evaluation Forum, http://www.clef-campaign.org
[4] http://en.wikipedia.org/wiki/Mean_reciprocal_rank
[5] http://sourceforge.net/projects/jirs/

model and the semantic resource used. With respect to the semantic resource level, the current release of AWN (Elkateb et al., 2006) allows to work with around 20,000 words grouped into 10,000 synsets.

At the experiments level, the current evaluation process uses a set of 82 CLEF questions 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.) as illustrated in Table 1. The considered questions are related to different topics and, therefore, present a significant coverage.

Domains	# Q	%
History	20	24,69
Sport	5	6,17
Politic	12	14,81
Culture	9	11,11
Geography	8	9,88
Technology	7	8,64
Other	21	25,93

Table 1. The distribution of the considered question per domain

Actually, we did not use any specific Q/A system. However, in order to simulate the use of our QE module within Q/A systems, we carried out the following experiment process: we manually select the most relevant keywords of each question. The given keywords are, then, extended using our semantic Query Expansion process. After that, the answer of each question and extended question is searched within the first five snippets returned by the used Passage Retrieval (PR) system. In the keyword-based evaluation process only a search engine (Yahoo[7]) is used as PR system (the returned snippets are considered in the evaluation).

In the structure-based evaluation process, JIRS is used. Indeed, JIRS is a language-independent PR system which has been already adapted to a few non-agglutinative European languages (such as English and French) as well as to the Arabic language (Benajiba et al., 2007b). The re-ranking of the retrieved passages is based on a distance density n-gram model. In (Benajiba et al., 2007b) authors explain the idea of this model which gives more weight to the passages where the most relevant question structures appear nearer to each other. In (Gomez et al., 2007) some experiments were carried out to re-rank snippets obtained with Yahoo in order to return the most relevant ones containing the answer.

In the next section we present the results obtained with an automatic evaluation process using the Yahoo search engine.

3 Keyword-based experiments

In this section, we investigate whether or not the semantic Query Expansion succeeded in improving results with respect to when no semantic QE is employed.

Type QE	Accuracy	MRR
Without Semantic QE	**1,22%**	**0,99**
QE by Synonyms	3,66%	1,63
QE by Definitions	4,88%	2,16
QE by Subtypes	4,88%	2,39
QE by Supertypes	8,54%	3,49
Overall Semantic QE	**7,32%**	**3,25**

Table 2. Experiment results of AAWN Query Expansion using Yahoo API

Table 2 shows the results of the experiments using the Yahoo API. The poor performance obtained is due to the fact that some relaxations are not used when we perform an automatic process. For example, in the manual process we can identify answers composed of more than one word. For example, the answer of the question **"** كم تبلغ القيمة المادية لجائزة المغرب للكتاب؟" (What is the value of the Moroccan book award?) is "7000 دولار". If a snippet contains for instance the expression " سبعة آلاف دولار" the answer is considered correct.

Nevertheless, even with an automatic process, the use of AAWN Query Expansion has improved the accuracy (from 1,22% to 7,32%) and the MRR (from 0,99 to 3,25). Moreover, the use of only one type of semantic QE already improves of the considered measures. For instance, the QE by synonym reaches an accuracy of 3,66 % (against 1,22% without QE) and 1,63 as MRR (against 0,99 without QE).
We can also notice that among the four semantic types of QE the one by supertypes gives the best results in term of accuracy and MRR as well. The ranking of the four semantic types of QE is the same with respect to the accuracy measure or the MRR.

[6] http://www.dsic.upv.es/grupos/nle/downloads.html
[7] www.yahoo.com

Table 3 shows the statistics related to the average of the number of generated keywords per question and per type of QE. It lists also the number of answered question per type of QE.

Type QE	Avg (keywords/Q)	#Answered Questions
By Synonyms	2,28	10
By Definitions	0,99	14
By subtypes	1,05	16
By supertypes	1,24	18
Semantic QE	5,56	18
Without QE	0	8

Table 3. Answered questions per type of QE using Yahoo API

For 21,95% of the considered questions the answer was found (in one of the first five snippets) after using the semantic QE. Without semantic QE we reach only a percentage of 9,76 %.

In Figure 2 we present graphically statistics related to the average of the number of generated keywords per question and per type of QE. Although the QE by synonyms generates an average of 2,28 keywords per question, the number of answered question using this semantic relation is the least among the four relations. Their ranking with respect to the number of answered question is the same as the accuracy and MRR measures.

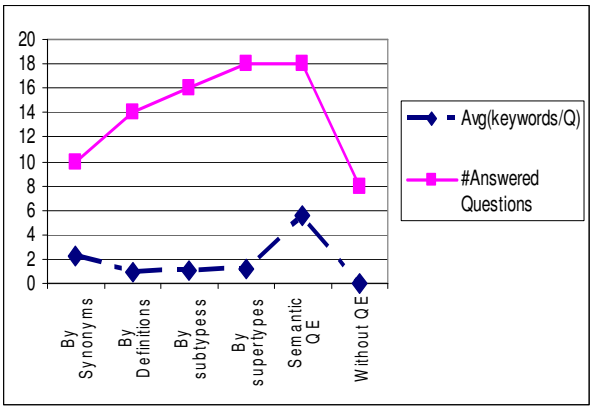

Figure 2. Answered questions per type of QE using the Yahoo API

In the next section we perform the same experiments using the JIRS PR system which considers the comparison of the returned passages structure with the one of the question.

4 Structure-based experiments

In the previous section we have considered only the first five snippets. However, the expected answer could exist in the returned results but not in these first five snippets. As we already mentioned, Gomez et al. (2007) have carried out preliminary experiments that show how JIRS PR system helps to re-rank Yahoo search engine snippets in order to make easier the answer extraction. Indeed, they have showed that the distance density n-gram model of JIRS improves both the coverage and the redundancy of the answers.

In these experiments, we have built a corpus based on the first 1,000 returned snippets by Yahoo. For the 82 CLEF questions, we have obtained an average of 42.96 returned snippets per question. After applying the JIRS indexation process to the built corpus, we have carried out the same experiments of the previous section, but using the JIRS PR system instead of the Yahoo API. Table 4 below shows the new results reached.

Type QE	Accuracy	MRR
Without Semantic QE	**15,85%**	**5,46**
QE by Synonyms	18,29%	6,72
QE by Definitions	9,76%	3,54
QE by Subtypes	8,54%	3,93
QE by Supertypes	13,41%	5,35
Overall Semantic QE	**19,51%**	**7,85**

Table 4: Experiment results of AAWN Query Expansion using the JIRS

Comparing the new results with the previous ones, the accuracy and the MRR have been improved even when we did not use any QE. The use of the proposed semantic QE together with JIRS improves the relevance of the first five snippets returned by the search engine. Therefore, the snippets are re-ranked better.

The results show that the proposed semantic Query Expansion continues to improve the accuracy and the MRR even if with different types of PR. The whole semantic QE has obtained an accuracy of 19,51% (against 15,85% without QE) and an MRR of 7,85 (against 5,46 without QE).

On the other hand, the QE by synonyms has provided the best results unlike the previous experiments. Indeed, the obtained accuracy is 18,29% and the MRR is 6,72. The ranking of the

QE types has changed, and the supertype-based QE provides, in this case, the second best performance instead of the first one.

In Table 5 and Figure 3 we show the statistics regarding the number of answered questions.

Type QE	Avg (keywords /Q)	#Answered Questions
QE by Synonym	2,28	22
QE by Definition	0,99	12
QE by subtypes	1,05	14
QE by supertypes	1,24	19
Overall Semantic QE	5,56	24
Without QE	0	23

Table 5. Answered question per type of QE using JIRS

The number of answered questions has passed to 24 (against 18 previously) in the case of semantic QE. However, the number of answered questions in the case of not using QE is approximately the same.

The ranking of the semantic QE types according to this measure confirms the one obtained with the accuracy and the MRR.

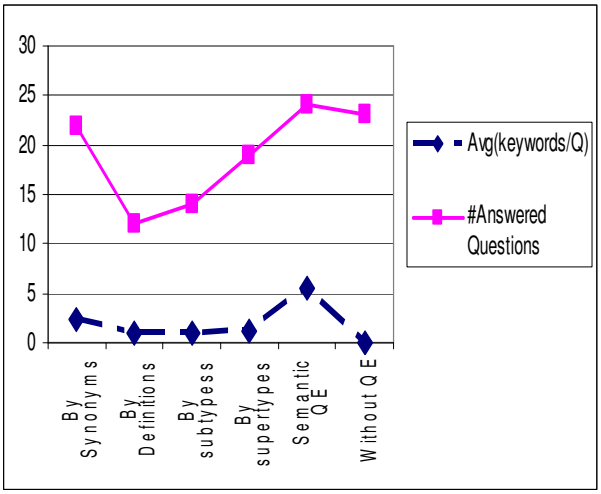

Figure 3. Answered questions per type of QE using the JIRS

Figure 3 shows that the number of answered questions does not clearly depend on the number of used keywords.

5 Synthesis of the evaluation results

In order to make a summarized comparison between the keyword-based experiment and the structure-based experiments, we have drawn the following tables (see table 6 and table 7):

Accuracy	Without JIRS	Using JIRS
Without Semantic QE	**1,22%**	15,85%
With Semantic QE	7,32%	**19,51%**

Table 6. Comparison of the Accuracy reached in the two experiments

MRR	Without JIRS	Using JIRS
Without Semantic QE	**0,99**	5,46
With Semantic QE	3,25	**7,85**

Table 7. Comparison of the MRR reached in the two experiments

Table 6 illustrates that the best accuracy and MRR (19,51% and 7,85) have been obtained when the semantic Query Expansion and the JIRS PR system are used together.

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 without using semantic QE.

6 Conclusion and Future Works

This work has been done in order to evaluate our proposed semantic QE for Arabic Q/A systems. Our aim was to confirm the preliminary experiments which showed that the accuracy and the MRR have been improved and that our semantic QE process (based on the current release of AWN) is adequate to improve the passage retrieval stage of an Arabic Q/A system.

This work has confirmed that, once more, the semantic QE improves both the accuracy and the MRR. In addition, in the case where it is combined with JIRS, our approach has obtained an accuracy around 19,51% and 7,85 as MRR. This means that when we take into account the semantic and the structure of the question we improve the probability of obtaining relevant passages (i.e., containing the answer).
The use of Arabic WordNet within the Amine Platform traces new ways regarding the semantic QE. As future work, we could take advantage of the concept definitions. Indeed, we could calcu-

late the similarity between the question and the returned passages according to a semantic comparison.

The proposed semantic QE approach does not define, so far, any weight to be assigned to the generated keywords. In the next steps of this work, we could decide on the relevance of each keyword according to its source (e.g. QE by supertypes could have the higher value) and the distance between the generated keyword and the initial one.

Finally, at the moment, the AWN project does not cover totally the standard Arabic. Therefore, the consideration of a completed version of AWN (Rodriguez et al., 2008) is to be intended.

Acknowledgement

This research was made possible thanks also to the following projects: AECI-PCI A010317/07, AECID PCI B017961/08 and CICYT TIN2006-15265-C06. We would like also to thank Dr. Manuel Gómez (Universidad Alicante, Spain) to help us with the use of JIRS.

References

Abouenour L., Bouzoubaa K., Rosso P. 2008. Improving Q/A Using Arabic Wordnet. In: Proc. *The 2008 International Arab Conference on Information Technology (ACIT'2008),* Tunisia, December.

Benajiba Y., Rosso P., Lyhyaoui A. 2007a. "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.

Benajiba Y., Rosso P., Gómez J.M. 2007b. "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.

Brini W., Ellouze M., Hadrich Belguith L. 2009. *QA-SAL :* "Un système de question-réponse dédié pour les questions factuelles en langue Arabe*". In: 9ème Journées Scientifiques desJeunes Chercheurs en Génie Electrique et Informatique, Tunisia.* (To appear in March; in French).

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

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

Fellbaum C. 2000. "WordNet: An Electronic Lexical Database". *MIT Press*, *cogsci.princeton.edu/~wn,* September 7.

Gómez J. M., Rosso P., Sanchis E. 2007. Re-ranking of Yahoo snippets with the JIRS Passage Retrieval system. In: Proc. Workshop on Cross Lingual Information Access, CLIA-2007, 20th Int. Joint Conf. on Artificial Intelligence, IJCAI-07, Hyderabad, India, January 6-12.

Hammou B., Abu-salem H., Lytinen S., Evens M. 2002. "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.

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

Mohammed F.A., Nasser K., Harb H.M. 1993. "A knowledge-based Arabic Question Answering System (AQAS)". *In: ACM SIGART Bulletin*, pp. 21-33.

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

Niles I., Pease A. 2003. "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.

Rachidi T., M. Bouzoubaa, L. ElMortaji, B. Boussouab, and A. Bensaid. 2003. "Arabic user search Query correction and expansion", in Proc. of COPSTIC'03, Rabat December 11-13.

Rodríguez H., Farwell D., Farreres J., Bertran M., Alkhalifa M., Antonia Martí M., Black W., Elkateb S., Kirk J., Pease A., Vossen P., and Fell-

baum C. 2008. Arabic WordNet: Current State and Future Extensions in: Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008. http://www.lsi.upc.edu/~nlp/papers/rodriguez08a.pdf.

Rosso P., Benajiba Y., Lyhyaoui A. 2006 "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.

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