An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering

Lahsen Abouenour, Karim Bouzouba, and Paolo Rosso

Abstract—**This paper describes an approach for improving the re-ranking of passages for Arabic Question/Answering (Q/A). This approach implements a process performing a semantic Query Expansion (QE) based on the Arabic WordNet (AWN) ontology with a structure-based Passage Retrieval (PR) based on the Distance Density n-gram model. Experiments with a set of translated CLEF and TREC questions have shown that the accuracy, the Mean Reciprocal Rank and the number of answered questions have been significantly improved using our approach. These experiments have been conducted in an open domain by querying the web as document collection. An analysis of the obtained performances is discussed in this paper.**

Index Terms—**Question/Answering; Semantic Query Expansion; Arabic WordNet; Distance Density n-gram model; JIRS**

I. Introduction

QUESTION/ANSWERING (Q/A) systems belong to the category of advanced Information Retrieval (IR) tools. They differ from the widely used Search Engines (SEs) since a precise answer is returned to the user rather than a list of snippets. Indeed, the use of SE presents a constraint for users as they have to manually filter a long set of returned documents.

Research in the field of Q/A has known significant progress for languages such as English, Spanish, French or Italian [12]. In the context of the Arabic language there are few attempts for building Q/A systems. This may be due to the particularities of the language (short vowels, absence of capital letters, complex morphology, etc.). The most well-known Arabic Q/A systems are:

- QARAB [24] 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[1], 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 [31] is knowledge-based and, therefore, extracts answers only from structured data and not from raw text (non structured text written in natural language).
- ArabiQA [9] is an Arabic Q/A prototype based on the Java Information Retrieval System (JIRS)[2] [8] 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 [7] and a set of patterns manually built for each type of question.
- QASAL [10] is a recent attempt for building an Arabic Q/A which process factoid questions (e.g. questions that have NE answers). Experiments have been conducted and showed that for a test data of 50 questions the system reached 67.65% as precision, 91% as recall and 72.85% as F-measure.

AQAS and QARAB offered the Arabic Natural Language Processing (NLP) research community the first prototypes of Arabic Q/A systems. However, AQAS processes only structured data whereas QARAB provided passages instead of precise answers. ArabiQA and QASAL target only factoid questions. Whereas the former integrates a NER system that has been evaluated and tested, the latter has been also tested but the two tests have used a lower number of questions. The use of all these systems in an open domain such as the web has not been tested.

The generic architecture of a Q/A system is in Figure 1. Its main modules are:
(i) *Question analysis and classification module*. In this module a question is analyzed in order to extract its keywords, identify the class of the question and the structure of the expected answer, form the query to be passed to the PR module, etc.

Manuscript received October 30, 2009. This work is supported in part by the TEXT-ENTREPRISE 2.0 TIN2009-13391-C04-03 and PCI-AECID C/026728/09 research projects.

L. ABOUENOUR is with the Mohammadia School of Engineers, Agdal, Rabat, Morocco. phone: (+212) 664 06 35 01; (e-mail: abouenour@yahoo.fr).
K. BOUZOUBA is with the Mohammadia School of Engineers, Agdal, Rabat, Morocco; (e-mail: karim.bouzoubaa@emi.ac.ma).
P. ROSSO is with the Natural Language Engineering Lab., ELiRF, Dpto. Sistemas Informáticos y Computación, Universidad Politécnica Valencia, Spain; (e-mail : prosso@dsic.upv.es).

[1] http://www.raya.com
[2] http://sourceforge.net/projects/jirs

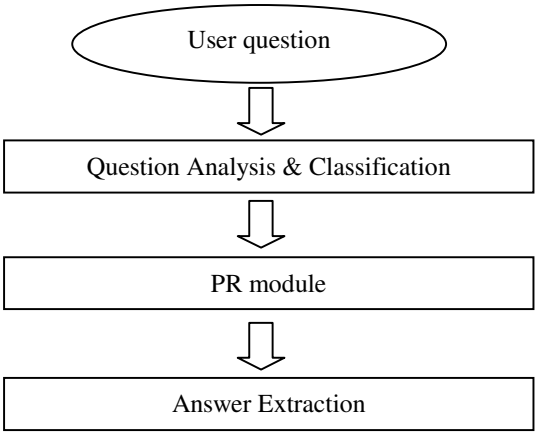

Fig. 1. The regular architecture of a Q/A system

(ii) *PR module*. This module is a core component of a Q/A system. The quality of the results returned by such system depends mainly on the quality of the PR module it uses [17]. Indeed, this module uses the query formed by the previous module and extracts a list of passages from an IR process (generally a SE such as Google[3] or Yahoo[4]). Thereafter, this module has to perform a ranking process in order to improve the relevance of the candidate passages according to the user question.

(iii) *Answer Extraction (AE) module*. This module tries to extract the answer from the candidate passages provided by the previous module. In advanced Q/A systems, this module can be designed to construct the answer from one or many passages. Of course, the AE module will fail to return the correct answer if the candidate passages provided by the PR module are not relevant and do not contain the answer.

The implementation of these three modules for the Arabic language is a challenging task. The few attempts, described previously, for building an Arabic Q/A system show that many efforts are still needed in order to reach the progress made in the same area for other languages. There are several lines that may be considered by researchers in this direction, namely: improving the question analysis task as well as the NER task for the purpose of Arabic Q/A systems, enhancing the PR module, adapting answer extraction techniques that have been evaluated for other languages in order to use them in the context of Arabic Q/A systems, etc.

The definition of the Q/A systems modules as described above shows the importance of the candidate passages generated by this module. Therefore, one of the possibilities to improve a Q/A system is to improve its PR module since its role is to get the most relevant passages from the documents with respect to the processed question. In order to do so, one promising way is to use instead of a keyword-based PR system (such as Google, Yahoo, etc) a structure-based PR system such as JIRS. The JIRS PR system is designed to improve passage re-ranking in the context of Q/A systems. In fact, it allows considering the question structure at the PR stage. The use of the language independent PR system JIRS has been evaluated for languages such as English, Spanish, Italian and French [12]. Moreover, by using JIRS, experiments have shown that the accuracy[5] and the coverage[6] are better than when we use a SE adapted to the PR task [21]-[22]. The obtained performances for the above languages are encouraging. A significant evaluation in the context of Arabic Q/A systems has to be done. At least, this evaluation will allow researchers to have an idea about the quality of the adaptation of JIRS to the Arabic language [8].

The use of a structure-based PR is not enough for the improvement of the PR module. Indeed, since there are many ways to formulate a question in natural language, a Query Expansion (QE) process can be used in order to overcome the situations where the PR process eliminates relevant passages containing other forms of the question keywords or words related to them. For instance, if the question contains the keyword طريق (Tryq : a way) the query used by the PR process can be expanded to include its other morphological forms like طرق (Trq : broken plural of Tryq) or طرقات (TrqAt : plural of Tryq). A more advanced QE process relies also on semantic relations. For example, we can include keywords like ممر (mmr : path) or مسار (msAr : trajectory) since they have a similar meaning to the original keyword. Some QE techniques using light-stemming can enhance recall[7] [23], while others improve precision[8] [2]. Generally, QE increases the recall at the expense of precision.

In the context of the Q/A task, the precision depends also on the question structure. Indeed, a document is relevant not only because it contains the question keywords (or expanded keywords) but also by containing words close to those of the question. For instance, let us consider the question **"** متى تم بناء قصر الحمراء **؟"** (When the Alhambra castle was built?): the keywords of this question are تم (tm : has been completed) – بناء(bnA' : building) – قصر (qSr : castle) – الحمراء (AlHmrA' : Alhambra). One of the most relevant passages should contain for instance the structure "...تم بناء قصر الحمراء...", "... تم تشييد قصر الحمراء..." or other similar structures. A passage which contains the structures "... تم بناء..." and "...قصر الحمراء..." separately is less relevant than the one containing "...تم تشييد قصر الحمراء...". The former contains all of the question keywords but in a

[3] http://www.google.com
[4] http://www.yahoo.com

[5] Accuracy is defined as the average of the questions where we find the answer in the first rank.
[6] The coverage gives the proportion of the question set for which a correct answer can be found within the top n snippets retrieved for each question.
[7] Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents (which should have been retrieved).
[8] Precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

different structure, whereas the latter, the most relevant, has the same structure of the question even with one expanded keyword.

The objective of our work is to contribute in the improvement of Arabic Q/A systems by enhancing the PR module. We propose two directions for such enhancement: firstly, a semantic QE is used in the aim to have a high level of completeness (recall) when the IR process retrieves passages; then a structure-based process is used for passage re-ranking in order to have the expected answer at the top of the candidate passages list. In the first task we have used the content and the semantic relations existing in the Arabic WordNet (AWN) ontology [16]. In the second task we have adopted the JIRS PR system which is based on the Distance Density n-gram model. This model finds question structures in the passages and gives a higher similarity value to those passages containing more grouped structures.

The evaluation process of our approach considers two of the most known measures in the context of Q/A systems: the accuracy and the Mean Reciprocal Rank[9]. This process uses a set of 1500 TREC[10] [38] and 764 CLEF[11] questions manually translated to the Arabic language. The obtained results show an improvement in the accuracy, the MRR and the number of answered questions.

The rest of the paper is structured as follows: Section II describes the semantic QE; Section III presents the structure-based technique adopted and its adaptation to the Arabic language; Section IV is devoted to the presentation of the experiments that we have conducted; in Section V we discuss the results of the experiments and section VI summarizes the main conclusions of this work.

II. Semantic Query Expansion for Arabic Q/A

QE is the process of adding a new list of terms to the user query in the context of an IR system [37]. In this context, potential documents satisfying the user query may not contain the keywords as they are formulated by the user, but keywords either differently formulated or having a close meaning. The new list of terms generated by a QE process would allow the system to consider these relevant documents.

Many QE techniques have been investigated by researchers in the IR field. The basic ones are those targeting to fix spelling errors by searching for the corrected form of the words [25]. Other QE processes rely on morphological relations and reformulate the user query by adding the different variations which are generated from keywords stems [27]. Although this QE technique produces higher recall [15]-[30], it is difficult to assert that it improves the precision. This is why researchers have investigated other QE techniques such as those using semantic relations. Generally, semantic QE process is performed by considering the synonyms of the query keywords. A thesaurus can be used as a base for such a process [32]. However, the use of a thesaurus, which is generally built on the basis of statistical techniques, presents many disadvantages. Indeed, building a thesaurus is a time consuming task since a great amount of data is to be processed. Moreover, the precision of thesaurus based QE in term of semantic distance has to be proved.

The use of an ontology rather than a thesaurus is another way to implement advanced semantic QE. While in thesauri only terms are related, ontologies introduce the notions of concepts and instances [28]-[40]. A concept describes an entity on an abstract level with generic properties, whereas an instance is an actual representation of this concept with specific values of these properties. In addition to the number of semantic relations existing in it, an ontology presents also the advantage of containing concept relations, of allowing semantic reasoning and cross language IR.

The adoption of ontologies raises the problem of the availability of those semantic resources especially for languages less rich in available resources such as Arabic. Nevertheless, the last decade has known a number of attempts aiming at offering electronic resources to NLP researchers.

Our semantic QE approach is based on the AWN[12] ontology [35]. This choice is due to the following advantages:

- The AWN ontology is a free resource for modern standard Arabic.
- It is based on the design and the content of Princeton WordNet (PWN) [18].
- AWN has a structure which is similar to wordnets existing for approximately 40 languages, including English, Italian, Spanish, French, Basque, Bulgarian, Estonian, Hebrew, Icelandic, Latvian, Persian, Romanian, Sanskrit, Tamil, Thai, Turkish, etc. Therefore, cross-language processes could be considered later as an enhancement of the present work.
- It is also connected to the Super Upper Merged Ontology (SUMO) [33]. Let us recall briefly that SUMO is an upper level ontology which provides definitions for general-purpose terms and acts as a foundation for more specific domain ontologies. It contains about 2000 concepts.

AWN offers the possibility to export its content and structure onto many formats so that researchers can use it in their context. Figure 2 illustrates the structure of AWN and its mapping onto the English WN.

[9] Mean Reciprocal Rank (MRR) is defined as the average of the reciprocal ranks of the results for a sample of queries (the reciprocal rank of a query response is the multiplicative inverse of the rank of the correct answer).

[10] Text REtrieval Conference, http://trec.nist.gov/data/qa.html

[11] Cross Language Evaluation Forum, http://www.clef-campaign.org

[12] http://www.globalwordnet.org/AWN/

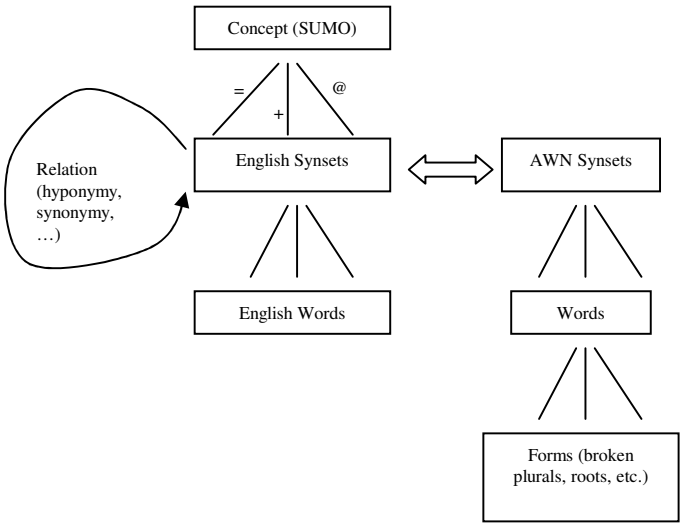

Fig. 2. The AWN data structure

As illustrated in Figure 2, the AWN data are divided into four entities:

- *Items* which are conceptual entities, including synsets (a set of words with the same part of speech that can be inter-changed in a certain context), ontology classes and instances. Besides a unique identifier, an item has descriptive information such as a gloss. Items lexicalized in different languages are distinct.
- *Word entity* is a word sense, where the citation form of the word is associated with an item via its identifier.
- A *form* is a special form that is considered dictionary information (not merely an inflectional variant). The forms of Arabic words that go in this entity are the root and/or the broken plural form, where applicable.
- A *link* relates two items, and has a type such as "equivalence," "subsuming," etc. Links connect sense items to other sense items, e.g. a PWN synset to an AWN synset, a synset to a SUMO concept, etc. Note that the "@", "+" and "=" symbols in the figure above refer to the INSTANCE_OF, MORE_GENERAL and EQUIVALENT mapping types respectively.

The current release of AWN contains: 11270 Arabic synsets (vs 115 000 synsets for English WN), 23496 Arabic words (vs 200 000 words for English WN). It contains also entries that are named entities (1142 synsets and 1648 words).

The AWN ontology contains different relations between its items such as hyperonymy/hyponymy (supertypes/subtypes relations), synonymy, meronymy/holonymy (part/whole relations), etc. Our semantic QE approach uses four semantic relations among those existing between AWN synsets (items), words and forms. Therefore, the approach distinguishes four sub QE processes: (i) *QE by synonyms*, (ii) *QE by definitions*, (iii) *QE by subtypes* and (iv) *QE by supertypes*. Unlike the QE by subtypes and supertypes, the QE by synonyms and definitions generate new terms which are not in the question term neighborhood. Moreover, our process is recursive generating other terms which are not reachable by the simple use of classical QE methods.

Let us consider the example of the question " ما هو المنصب الذي تقلده سيلفيو برلسكوني؟" (e.g. What position does Silvio Berlusconi hold?). For this question, the Google SE returns the first five snippets listed in Table I. Even though these snippets contain the keywords "سيلفيو برلسكوني" (Silvio Berlusconi) and "منصب" (manoSib : position) there is a need to identify also the related terms to these keywords. Indeed, at the answer extraction stage we have to find units of text containing the (or a similar) structure to the expected answer. In the given example the expected answer is something like "[Answer] تقلد سيلفيو برلسكوني" where [Answer] is any term of group of terms that is semantically related to the keyword "منصب" (manoSib : position).

The idea is to apply our semantic QE process in a way to have new terms related to the considered keyword. After that, we re-rank the passages in order to have in the first ranks those containing both the question keywords and the new generated terms close each to another.

Our QE process is applied only for keywords which are not stopwords, namely: ما (mA : what), هو (hw : he) and الذي (Al*y : that). For instance, the term "المنصب" (manoSib : position) belongs to the AWN synset "مَنْصِب - وَظِيفَة" (manoSib : position – wZyfp : job).

TABLE I
SNIPPETS RETURNED BY THE GOOGLE SE FOR THE QUESTION
"ما هو المنصب الذي تقلده سيلفيو برلسكوني؟"

Snippet ID	Snippet
1	3 أيلول (سبتمبر) 2009 **...** تسود أرجاء ايطاليا الآن دوي صوت رئيس وزرائها **سيلفيو برلسكوني** في حديثه الذي **...** وفساد ففي عام2004 وخلال **تقلده المنصب**, برأت محكمة ايطالية **برلسكوني ...** الوزراء الايطالي من الهجوم الذي يتعرض له, لأن كل ما يهم هو تصرفات **...**
2	وفي روما وجه رئيس **الحكومة** الإيطالية **سيلفيو برلسكوني** رسالة تعزية شخصية إلى **....** ووصف رئيس لجنة الأنموفيك السابق الذي سيغادر **منصبه** نهاية هذا الشهر بعض أعضاء **...** وقياسا على ما هو مألوف بالنسبة للأكاديمي العادي الذي ينفق حياته في كتابة عدد من **.... تقلد مناصب** سياسية وعسكرية فكان وزيرا أول ثم قائد أركان الجيش الوطني. **...**
3	29 آب (أغسطس) 2005 **...** أعترف أن رئيس الوزراء الإيطالي، رجل الأعمال البارز، المثير للجدل، **سيلفو برلسكوني، شخصية** محببة لي. فأنت إن اتفقت مع هذا الرجل الذي يبلغ من **...**
4	**نظام** الحكم جمهوري برلماني رأس الدولة الرئيس جورجو نابوليتانو **...** عرفت إيطاليا تقلب **سياسي** منذ العهد السياسي الجديد بعد نهاية الحرب العالمية **...** ميدياست، اللتي يملك رئيس الوزراء الإيطالي **سيلفيو برلسكوني** معظم أسهمها...
5	الاضواء مسلطة/ محكمة إيطالية تنزع الحصانة عن **سيلفيو برلسكوني .....** الاقتصادية/ نحو تريليون دولار يتعين ان يتم انفاقها خلال 18 شهرا لخلق **وظائف** وضرائب اقل...

In addition to its synonym "wZyfp : job" the considered entry has also two direct supertypes in AWN which are مهنة (mhnp : job) and نشاط (n$AT : activity). However, it has no subtypes. The SUMO concept related to the considered synset

is "POSITION". The definition of this concept in the SUMO ontology is as follows: "A formal position of responsibility within an &%Organization. Examples of Positions include president, laboratory director, senior researcher, sales representative, etc.". Given that the SUMO concepts are preceded by the symbols "&%" and "?", we can identify the SUMO concept "ORGANIZATION" as being related to the "POSITION" concept. This new concept is linked to the AWN synset presented by the term "جَمْعِيَّة" (jamoEiy~ap : association). The neighborhood (supertypes and subtypes) of this new synset allows us to reach new terms such as: "مُنَظَّمَة" (munaZ~amap : organization), "جَمَاعَة" (jamaAEap : community), "حكومة" (Hkwmp : government) and "نِظَام سِيَاسِي" (niZaAm siyaAsiy : political system). The SUMO concept "ORGANIZATION" is also linked to the synset represented by the term "رَئِيس" (ra}iys : Chairman). New terms could be reached in the neighborhood of this synset such as مَلِك (malik : king), رَئِيس الوِزَرَاء (ra}iys AlwizaraA' : prime minister) and رَئِيس الدَّوْلَة (ra}iys Ald~awolap : head of nation). Figure 3 illustrates the result of the recursive QE process that we have performed starting from the question keyword "المنصب". Note that boxes with labels 1, 2, 3 and 4 refer respectively to the QE by synonyms, definition, subtypes and supertypes. Note that the non expanded boxes refer to a non existing AWN entry (synonym, definition, subtype or supertype).

Our QE process generates three groups of new terms:

- Terms reached by the hyponymy (subtypes) and hypernymy (supertypes) relations: "مهنة" (mhnp : profession), "تَفَاوَضَ" (tafaAwaDa : negotiation), "قِيَادَة" (qiyaAdap : command), "ضَبْط" (DaboT : control), "صَنْعَة" (SanoEap : workmanship), "عَمَل" (Eamal : work) and "نشاط" (n$AT : activity). These terms represent the direct neighborhood of the given question keyword.
- Terms such as رَئِيس (ra}iys : president) and جَمْعِيَّة (jamoEiy~ap : association) which do not exist in the direct neighborhood of the considered AWN synset but can be reached through the definition of the SUMO concept equivalent to that synset.
- Terms which do not exist in the direct neighborhood of the considered AWN synset but can be reached through the SUMO concept "IntentionalProcess" equivalent to the second supertype of the given synset. The definition of this concept uses two other SUMO concepts: "CognitiveAgent" and "Process". The former is equivalent to the synset represented by the terms "شَخْصِيَّة" ($axoSiy~ap : personality) and "ذات" (aAt* : self). The latter is equivalent to the synset symbolized by "حَدَث" (Hadav : evant) and "وُقُوع" (wuquwE : occurring).

Using our QE process we have reached new terms which are semantically related to the question keyword "المنصب". The returned snippets using the expanded question (e.g. the user question where the keyword "المنصب" is replaced with each generated keyword) are more relevant since they will contain terms such as "حكومة" (Hkwmp : government) and رَئِيس الوِزَرَاء (ra}iys AlwizaraA' : prime minister). As Table I shows the expected answers "رئيس الوزراء الإيطالي " (the Italian prime minister) or "رئيس الحكومة الإيطالية" (the president of the Italian government) exist in the returned snippets and are reached only through the terms that have been generated by our semantic QE. Other examples showing the usefulness of using this semantic QE process could be found in [3].

Our process will also generate irrelevant or less relevant terms which, passed to the SE, will produce irrelevant snippets. Moreover, a bootstrapping method can lead to an indefinite QE process or at least can increase the number of irrelevant snippets retrieved as a result of the expanded query. Thus, a threshold is to be set in order to avoid such undesired behavior. Preliminary experiments that we have conducted [4] show that we can for instance set a two-level threshold for QE by subtypes and supertypes as this threshold helps in improving performances without producing a great amount of new terms. Since the current AWN release is limited in term of coverage and available links between synsets and SUMO concepts, we did not set a threshold for the QE by synonyms and definitions. The use of AWN as described above generates a significant amount of new terms that can be used in the query passed to the PR process.

The preliminary experiments [5]-[6] allowed us to evaluate the improvement of the accuracy and the MRR when the semantic QE is used for 82 CLEF and 82 TREC questions. Table II below summarizes the obtained results.

The accuracy and the MRR have both been improved when using our QE process. For instance, using QE with CLEF questions we have obtained 7.32% (1.22% without QE) as accuracy and 3.25 (0.99 without QE) as MRR.

TABLE II
RESULTS OF THE PRELIMINARY EXPERIMENTS REGARDING THE USE OF THE SEMANTIC QE

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	1.22%	**7.32%**	5.02 %	**6.95%**
MRR	0.99	**3.25%**	2.04	**2.88%**

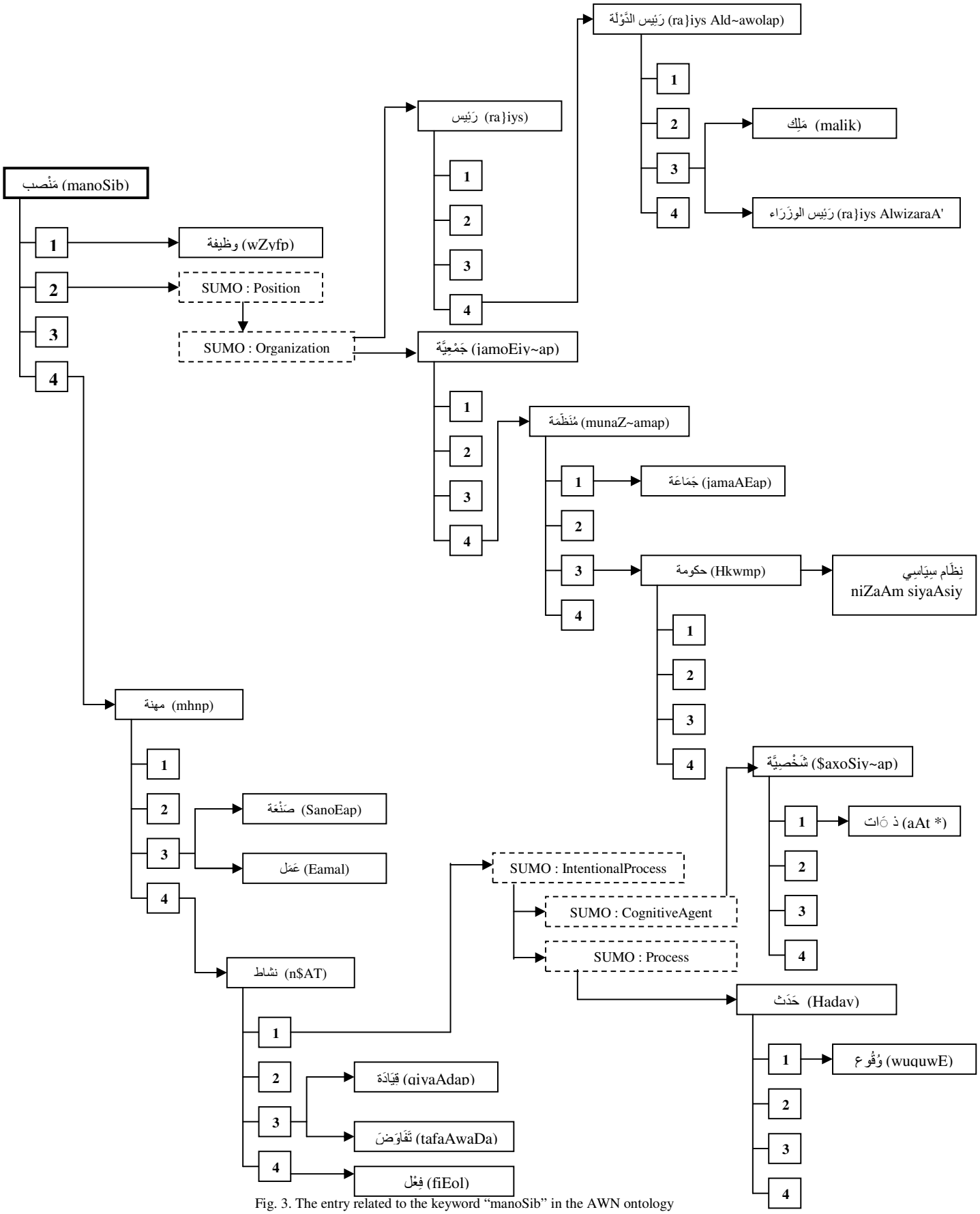

Fig. 3. The entry related to the keyword "manoSib" in the AWN ontology

Let us recall that our QE process aims to reach a high level of completeness. The retrieved passages using this process have to be re-ranked later using a structure-based approach. The Distance Density n-gram is a model designed for passage re-ranking with respect to the similarity between a retrieved passage and the original user question. In the next section, we move to the presentation of the structure-based passage re-ranking using the Distance Density n-gram model. We also describe the JIRS system which implements this model.

III. STRUCTURE-BASED PASSAGE RE-RANKING

The most natural manner to present the results of an IR system to the user is showing the most relevant documents at the top of the list of the results. In order to be able to do so, it is necessary to first rank the obtained documents according to their relevance before displaying them to the user. Therefore, a ranking process is to be performed in order to assign higher weights to those documents which better match the user query. This process can be based on passages instead of documents. A PR based ranking process tries to provide relevant units of text related to the user query. PR ranking compare passages having the same length rather than documents with a different length.

There are different methods to perform the segmentation of documents into passages [36]:

- Dividing the text by considering the units according to their semantic meaning and topics change;
- Using the explicit structure of the documents e.g passages can be extracted from the tags of an SGML document;
- Considering parts of the text containing a fixed number of words;
- Getting arbitrary passages [26];
- etc.

Recent PR approaches use statistical models of documents and queries ("language models") in the context of IR [39]-[29]-[34]. In the present work, we have been interested in the adoption of the Distance Density N-gram model since it presents the advantage of being tested through different experiments [14]-[1]-[19]. Indeed, these works have proved that the density approach is the most successful technique especially for the Q/A systems. The Java Information Retrieval System (JIRS) is a language independent PR system, implemented on the basis of the density model, that has been adapted to work also with the Arabic language [8]. Before presenting the features of the JIRS PR system, we give a brief description of how the density model could improve the passage ranking process. Let us recall that this process is based on the assignment of weights to the retrieved passages. The density model is designed in a manner to give more weight to those passages where the question terms appear nearer to each other. In order to implement such a model, two steps are needed. In the first step, passages are searched and assigned a weight which is expressed as:

$$w_k = 1 - \frac{\log(n_k)}{1 + \log(N)} \quad (1)$$

Where n_k is the number of passages in which the term associated with the weight w_k appears and N is the number of the considered passages.

The second step uses a model which gives more importance to passages where the question n-grams present a higher density. This model can be expressed as:

$$Sim(p,q) = \frac{1}{\sum_{i=1}^{n} w_i} \cdot \sum_{\forall x \in P} h(x) \frac{1}{d(x, x_{max})} \quad (2)$$

Where x is an n-gram of p formed by q terms, w_i are the weights defined by (1), $h(x)$ can be defined as:

$$h(x) = \sum_{k=1}^{j} w_k \quad (3)$$

$d(x,xmax)$ is the factor which expresses the distance between the n-gram x and the n-gram with the maximum weight $xmax$, this factor is expressed by the formula:

$$d(x, x_{max}) = 1 + \ln(1 + L) \quad (4)$$

where L is the number of terms (including stopwords) between the n-grams.

The JIRS implements the density model described above. JIRS extracts the N-grams from the question and compares them with the N-grams extracted from the ranked passages returned by a SE such as Google, Yahoo or Lucene[13]. The final result of the system is a list of passages re-ranked with respect to the similarity of structure between them and the user question.

The system was reported to offer high performance in all of the Spanish, French and Italian languages [13]. The Arabic-JIRS version of the passage retrieval system relied on the same architecture as for the other languages. The main modifications were made on the Arabic language-related files (text encoding, stop-words, list of characters for text normalization, Arabic special characters, question words, etc.) [8].

In order to show the usefulness for using JIRS, let us consider the example of the previous section. Let us recall that the QE process that we have applied to the keyword "منصب" (manoSib : position) has generated new related terms like "حكومة" (Hkwmp : government), "نِظَام سِيَاسِي" (niZaAm siyaAsiy : political system), قِيَادَة (qiyaAdap : command) and

[13] http://lucene.apache.org/java/docs/

رَئِيس الوزَرَاء (ra}iys AlwizaraA' : prime minister). Using the original user question, JIRS returns the results listed in Table III below. Note that the collection used is created from the content of the first thirty snippets returned by the SE after querying it using the same question.

TABLE III
PASSAGES RETURNED BY JIRS FOR THE QUESTION
"ما هو المنصب الذي تقلده سيلفيو برلسكوني؟"

Passage	Similarity	Doc	Passage
2	0.44735444	1	تسود أرجاء ايطاليا الآن دوي صوت_رئيس وزرائها سيلفيو برلسكوني في حديثه الذي .. .وفساد ففي عام2004 وخلال تقلده المنصب, برأت محكمة ايطالية برلسكوني .. .الوزراء الايطالي من الهجوم الذي يتعرض له, لأن كل ما يهم هو تصرفات. ..
1	0.433959	1	3أيلول (سبتمبر) 2009 .. .تسود أرجاء ايطاليا الآن دوي صوت رئيس وزرائها سيلفيو برلسكوني في حديثه الذي .. .وفساد ففي عام2004 وخلال تقلده المنصب, برأت محكمة ايطالية برلسكوني. ..
4	0.35973868	2	وفي روما وجه رئيس الحكومة الإيطالية سيلفيو برلسكوني رسالة تعزية شخصية إلى ... .ووصف رئيس لجنة الأنموفيك السابق الذي سيغادر منصبه نهاية هذا الشهر بعض أعضاء. ..
13	0.35973868	4	عرفت إيطاليا تقلب سياسي منذ العهد السياسي الجديد بعد نهاية الحرب العالمية .. .ميدياست، اللتي يملك رئيس الوزراء الإيطالي سيلفيو برلسكوني معظم أسهمها. ..
14	0.35973868	5	الاضواء مسلطة/ محكمة إيطالية تنزع الحصانة عن سيلفيو برلسكوني ... .الاقتصادية/ نحو تريليون دولار يتعين ان يتم انفاقها خلال 18 شهرا لخلق وظائف وضرائب اقل. ..
9	0.16316938	3	29آب (أغسطس) 2005 .. .أعترف أن رئيس الوزراء الإيطالي، رجل الأعمال البارز، المثير للجدل، سيلفو برلسكوني، شخصية محببة لي .فأنت إن اتفقت مع هذا الرجل الذي يبلغ من. ..

As Table III shows, passage 2 has been assigned the best similarity score calculated by JIRS according to the Distance Density model. Even if passages 4 and 13 contain two structures closely similar to the one of the question. This is due to the fact that passage 2 contains two sub structures " سيلفيو برلسكوني" and "المنصب" which exist in the question, whereas passages 4 and 13 contain only one sub structure "سيلفيو برلسكوني". At the answer extraction level it would be easy to extract " رئيس الوزراء الإيطالي" and "رئيس الحكومة الإيطالية" as answers if passages 4 and 13 have been assigned a best score.

Let us now see how JIRS combined with the results of our QE process will improve passage ranking. The idea is to form expanded questions by replacing the keyword "المنصب" in the question with each term generated by our QE process. After that we use the expanded questions as queries passed to JIRS. Therefore, we have a ranked list of passages for each query. In our process we consider the best scored passages among the different queries. Table IV shows the list of ranked passages.

Table IV illustrates the added value of using JIRS and semantic QE together. Indeed, passages 13 and 4, which contain the expected answer, are now assigned a higher similarity scores.

TABLE IV
TOP RANKED PASSAGES RETURNED BY JIRS FOR THE EXPANDED QUESTIONS

Passage	Similarity	Doc	Passage
13	0.5583812	4	عرفت إيطاليا تقلب سياسي منذ العهد السياسي الجديد بعد نهاية الحرب العالمية .. .ميدياست، اللتي يملك رئيس الوزراء الإيطالي سيلفيو برلسكوني معظم أسهمها .. .
4	0.5294399	2	وفي روما وجه **رئيس الحكومة الإيطالية سيلفيو برلسكوني** رسالة تعزية شخصية إلى ... .ووصف رئيس لجنة الأنموفيك السابق الذي سيغادر منصبه نهاية هذا الشهر بعض أعضاء .. .
1	0.45635238	1	3 أيلول (سبتمبر) 2009 .. .تسود أرجاء ايطاليا الآن دوي صوت **رئيس وزرائها** سيلفيو برلسكوني في حديثه الذي .. .وفساد ففي عام2004 وخلال تقلده المنصب, برأت محكمة ايطالية برلسكوني .. .
2	0.44735444	1	تسود أرجاء ايطاليا الآن دوي صوت **رئيس وزرائها** سيلفيو برلسكوني في حديثه الذي .. .وفساد ففي عام2004 وخلال تقلده المنصب, برأت محكمة ايطالية برلسكوني .. .الوزراء الايطالي من الهجوم الذي يتعرض له, لأن كل ما يهم هو تصرفات .. .
1	0.433959	1	3 أيلول (سبتمبر) 2009 .. .تسود أرجاء ايطاليا الآن دوي صوت **رئيس وزرائها** سيلفيو برلسكوني في حديثه الذي .. .وفساد ففي عام2004 وخلال تقلده المنصب, برأت محكمة ايطالية برلسكوني .. .
13	0.5583812	4	عرفت إيطاليا تقلب سياسي منذ العهد السياسي الجديد بعد نهاية الحرب العالمية .. .ميدياست، اللتي يملك **رئيس الوزراء الإيطالي** سيلفيو برلسكوني معظم أسهمها .. .

In order to show the usefulness of the proposed approach which combines the semantic QE with JIRS, we conducted preliminary experiments on 82 CLEF and 82 TREC questions [6]. Table V below summarizes the obtained results.

TABLE V
RESULTS OF THE PRELIMINARY EXPERIMENTS REGARDING THE USE OF THE SEMANTIC QE TOGETHER WITH JIRS

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	15.85%	**19.51%**	2.7 %	**10.81 %**
MRR	5.46	**7.85**	0.67	**4.53**

The use of our QE based on AWN together with the structure-based re-ranking using JIRS gives the best performances in terms of accuracy (19.51% and 10.81%) and MRR (7.85 and 4.53).

In the current work, and in order to make significant conclusions, we have conducted new experiments using our semi-automatic built question set. This set contains not only 164 questions (82 from CLEF and 82 from TREC) as in the preliminary experiments, but almost all available CLEF and TREC questions (2,264 questions) translated to the Arabic

language. The next section is devoted to the presentation of the obtained results.

IV. Experimental results

A. Data Set

In the Q/A field, researchers have two well-known international competitions where they can compare their systems: the TREC and CLEF. In these competitions, works related to both monolingual and cross-lingual QA tasks are addressed. The test data provided by the two competitions cover a considerable variety of languages (English, French, Spanish, Italian, Dutch, etc.). Unfortunately, the Arabic language is not among them. Therefore, a need of a translation into the Arabic language of the available data set is to be done. In the context of the current work, we have manually translated all the TREC and CLEF questions available in English and French. Using these two test data sets allows us to conduct experiments with the same distribution of questions in terms of covered topics, question categories, nature of the expected answer, etc.

The numbers of translated questions[14] are: 1500 for the TREC set and 764 for the CLEF set. These questions are classified into different domains (sport, geography, politic, etc.) and different types. The types are defined from the nature of the expected answer. The considered types are:

- **MEASURE**: for instance "What distance does the Granada-Dakar rally cover?" " ما هي المسافة التي يغطيها رالي غرناطة - دكار؟"
- **ABREVIATION**: for instance "What is NASA?" " ما هي ناسا ؟"
- **COUNT**: for example "How many people are killed by landmines every year?" " كم عدد الاشخاص الذين يقتلون سنويا من جراء الألغام الأرضية ؟"
- **PERSON**: "What is the name of the Queen of the Netherlands?" " ما هو اسم ملكة هولندا ؟"
- **OBJECT**: "What is exhibited in the Vitra Design Museum?" " ما الذي يعرض في متحف فيترا للتصميم ؟"
- **LOCATION**: for instance "What is the capital of Chechnya?" " ما هي عاصمة الشيشان ؟"
- **ORGANIZATION**: "Which organization does Vanessa Redgrave support?" " ما هي المنظمة التي تدعمها فانيسا ريدجريف ؟"
- **TIME**: "When was the Universal Declaration of Human Rights approved?" " متى تمت المصادقة على الإعلان العالمي لحقوق الإنسان ؟"
- **LIST**: "Tell me names of robots." " أعطي أسماء روبوت."
- **OTHER**: " What is a risk factor for cardiovascular diseases? " " ما هو أحد عوامل الخطر لأمراض القلب والأوعية الدموية ؟"

Tables VI and VII below show, for each set, the number of questions belonging to the different question types.

TABLE VI
CLEF Questions per types

TYPE	#Q	%
PERSON	183	24%
LOCATION	123	16%
OTHER	115	15%
TIME	93	12%
COUNT	89	12%
ORGANIZATION	63	8%
ABREVIATION	34	4%
OBJECT	26	3%
MEASURE	16	2%
PERSON	183	24%

TABLE VII
TREC Questions per types

TYPE	#Q	%
OTHER	340	22,67%
LOCATION	307	20,47%
PERSON	258	17,20%
TIME	208	13,87%
ABREVIATION	133	8,87%
COUNT	106	7,07%
ORGANIZATION	57	3,80%
MEASURE	56	3,73%
OBJECT	29	1,93%
LIST	6	0,40%

As Table VI shows, the CLEF questions belong mainly to types which are NE. Indeed, roughly 60% of the questions concern PERSON, ORGANIZATION, TIME and LOCATION answers.

Table VII above shows that the major part of the TREC questions belongs to the types: LOCATION, PERSON and OTHER. The percentage of questions that are of NEs types is 55%.

B. Evaluation Process and Measures

The scope of our work is the improvement of passage ranking at the PR level for Arabic Q/A. An Arabic Q/A system is needed in order to embed a PR re-ranking process. For the purpose of the current work we just simulate a Q/A system. Therefore, our evaluation process is composed of automatic and manual tasks. Figure 4 below illustrates this process.

[14] available for download from www.emi.ac.ma/bouzoubaa/download.htm or from http://www.dsic.upv.es/grupos/nle/downloads.html

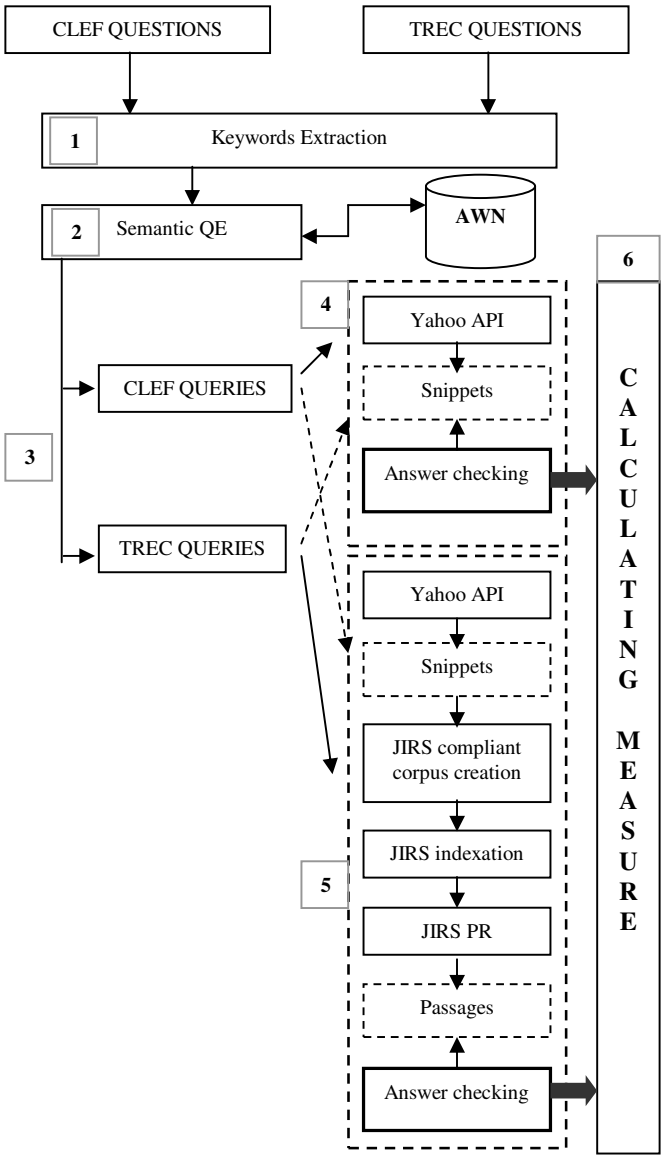

Fig. 4. The evaluation process steps

As illustrated in the figure above, each set of the considered questions is processed through the following steps:

- Step 1: the first step is to extract the relevant keywords contained in the question. The extracted keywords are to be passed to the QE process. We eliminate stopwords (a list of Arabic stopwords related to the Q/A task is available in JIRS [8]) in order not to consider them in the next step.
- Step 2: for each extracted keyword we perform the semantic QE process described previously in Section 2 (using AWN). Therefore, a list of new related terms is generated.
- Step 3: for each question set (e.g. TREC and CLEF) we generate a set of queries. Indeed, these queries are formed by replacing keywords in each question by their related terms generated in step 2.
- Steps 4 and 5: for each query set two types of experiments are conducted: Keyword-based and Structure-based experiments. In the first experiment we get the first five snippets returned by the Yahoo API for each query. After that, for each query, we check the existence of the corresponding answer in these snippets. A value $V_{k,j}$ is assigned to each question as follows:

$$V_{k,j} = \begin{cases} \mathbf{1} & \text{if the answer to question k has been found in the passage having the rank j (j is between 1 and 5)} \\ \mathbf{0} & \text{else} \end{cases}$$

In the structure-based experiment, we get m snippets returned by the Yahoo API in response to all the CLEF and TREC queries. We create then a collection of documents from the content of these snippets with respect to the format supported by JIRS. This collection is indexed using the corresponding JIRS process. After that, we make use of the monolingual retrieval process offered by JIRS in order to query our collection. We consider the first five results like what we have done in the keyword-based evaluation. In previous experiments it has been checked that the optimal value of m is between 800 and 1000 for the Spanish CLEF document collection [20]. For the purpose of our experiments we investigate the performances for m=200 in the case of the TREC questions and for m=1000 for the TREC and CLEF questions.

- Step 6: after each experiment, our process calculates the measures for the different set of questions. We have considered three measures:
  - The Accuracy, calculated according to the formula:

(5) $$\text{Acc} = \frac{1}{\text{Ns}} \sum_{k \in s} V_{k,1}$$

Where Ns is the number of questions of the question set s.

  - The Mean Reciprocal Rank, calculated as follows:

(6) $$MRR = Avg_{k \in s}\left(\frac{1}{5}\sum_{j=1}^{5}\frac{V_{k,j}}{j}\right)$$

Where k is a question belonging to the set s (CLEF or TREC), j is the passage rank.

  - The number of Answered Questions (AQ), the number of questions we find the answer in at least one of the first five ranks, is calculated according to the formula:

(7) $$AQ = \frac{1}{Ns}\sum_{k \in s}\max(Vk, j)$$

Where k is a question belonging to the set s, N is the number of question contained in the set s and $V_{k,j}$ the value assigned to the five passages returned in response to the question k.

C. Results

1) Query Expansion

For the TREC questions, the semantic QE has been performed for 858 questions. This means that AWN contains corresponding entries for 57.2% of the TREC questions. This percentage is higher in the case of the CLEF questions and reach 80.10% (612 questions out of 764). The overall coverage of AWN with respect to the two question sets is 64.93%. Table VIII and IX show the AWN coverage for the two question sets with respect to semantic relation type.

TABLE VIII
AWN SEMANTIC RELATIONS COVERAGE FOR THE TREC QUESTIONS

RELATION TYPE	#Q	%
Synonyms	850	99.07
Supertypes	179	20.86
Subtypes	132	15.38
Definitions	26	3.03

TABLE IX
AWN SEMANTIC RELATIONS COVERAGE FOR THE CLEF QUESTIONS

RELATION TYPE	#Q	%
Synonyms	608	99.35
Supertypes	143	23.37
Subtypes	102	16.67
Definitions	36	5.88

As Table IX shows, the coverage for the two question sets has the same trend in the four considered semantic relations. Indeed, for 99.35% of the questions there is at least one keyword that can be expanded by its synonyms in the AWN. So, for almost all the questions, at least one query can be formed by replacing the keyword in the question by one of its synonyms. However, the average of generated queries from the synonymy relation has been calculated and does not exceed 3.65 per question for the CLEF set and 4.26 for the TREC set. The coverage of AWN in terms of hyponymy (subtypes) and hypernymy (supertypes) relations does not exceed 25%. For the definition relation, the coverage is very low and is close to 6%.

The statistics above show that AWN is more developed with respect to the synonymy relation. However, efforts must be provided regarding the hierarchy of types and the connection with the SUMO ontology. In order to compare the performances before using our semantic QE and after, we consider only the subset of questions that can be expanded.

2) Keyword-based Evaluation

The aim of the keyword-based evaluation is to measure the usefulness of the semantic QE process in the context of the Q/A PR module. In this experiment, the passage ranking does not take into account the question structure. Table X shows the obtained results for the CLEF and TREC sets.

TABLE X
KEYWORD-BASED PERFORMANCES USING SEMANTIC QE FOR THE CLEF AND TREC QUESTIONS

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	5.07%	**8.35%**	3.38%	**5.24%**
MRR*	1.66	**3.12**	1.21	**2.04**
AQ	12.09%	**17.97%**	7.58%	**12.82%**

* Note that MRR has been multiplied by 100 in order to have a better readability.

From the table above, we can state that using our semantic QE improves the Accuracy, the MRR and the number of answered questions. Indeed, by using the QE based on AWN we have obtained 3.28% of gain, 1.46 and 5.88%, respectively, for the CLEF set and 1.86%, 0.83 and 5.24%, respectively, for the TREC questions. However, these performances are still lower compared to what we have reached in a previous work [4]. The particularity of that work is the fact that we manually checked the existence of the answer in the snippets returned by the Google SE. The accuracy reached was close to 33% while the MRR was around 10.

Indeed, many causes may be behind the failure of our process in identifying answers despite they exist in the first five passages. Generally, this failure is due to the multi-word answers, e.g. answers with more than one word. For instance, if the question is "متى ولد توماس مان ؟" (When was Tomas Mann born?) and the answer is "6 يونيو 1875" (6th June 1875) our process fails to get the answer in a passage containing the month and the year or the year only. Therefore, some relaxations are needed at the answer checking of our evaluation process in order to get the performances which are close to the reality.

We have considered three types of relaxations:

- For the date answers, if the process fails to get them we try then to search only the year.
- In the date answers, we search also with the Arabic corresponding months such as "أيلول" (September).
- If the process fails to identify the answer in a passage, we try identifying its stem instead of the entire word. We have used for so the Buckwalter morphological analyzer [11].

In addition to those relaxations, we perform, for multi-word answers, a subprocess which allows identifying passages that contains at least one word of the answer. For instance, the question "من الذي اخترع الهاتف ؟" (Who invented the telephone?) has the answer "الكسندر غراهام بيل" (Alexander Graham Bill), so if the word "الكسندر" (Alexander) appears in a passage then it is listed for a manual validation. Therefore, we obtain a list where each row contains the query, its answer, and the corresponding passage. This list is then manually checked in order to confirm that the passages contain the entire answers.

The manual validation allows us to avoid any impact that the mentioned relaxations could have on the results. Indeed, the aim behind using these relaxations is to downsize the number of passages in which correct answers are manually identified.

After considering these relaxations, we have obtained the results listed in Table XI.

TABLE XI
KEYWORD-BASED PERFORMANCES USING SEMANTIC QE FOR THE CLEF AND TREC QUESTIONS (AFTER RELAXATIONS)

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	11.76%	**14.40%**	8.16%	**12.35%**
MRR	3.85	**5.59**	3.1	**5.05**
AQ	25.16%	**29.74%**	16.78%	**23.43%**

The results presented in Table XI above show that the performances in term of accuracy, the MRR and the number of answered questions have been improved after considering the relaxations described previously.

Even if the use of our semantic QE improves the results, the reached performances are still unsatisfying. Let us now see what performances we can obtain by using a structure-based approach.

3) Structure-based Evaluation

The aim of the structure-based evaluation is to measure how JIRS can improve the relevance of the first five passages. A previous work has shown that JIRS improves the Yahoo snippets re-ranking for the English language [21]. In this experiment, we first evaluate whether or not JIRS improves this re-ranking in the context of the Arabic language. Secondly, we evaluate how our semantic QE combined with JIRS could reach higher performances in terms of accuracy, MRR and the number of answered questions. Tables XII and XIII show the performances reached after using JIRS for the two question sets. The results are presented before using the QE and after using it.

TABLE XII
STRUCTURE-BASED PERFORMANCES USING JIRS FOR THE CLEF AND TREC QUESTIONS (BEFORE RELAXATIONS)

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	8.77 %	**11.60%**	6.41%	**8.51%**
MRR	3.99	**5.26**	2.78	**3.7**
AQ	12.09 %	**16.01%**	6.99%	**10.60%**

TABLE XIII
STRUCTURE-BASED PERFORMANCES USING JIRS FOR THE CLEF AND TREC QUESTIONS (AFTER RELAXATIONS)

	CLEF		TREC	
MEASURES	Without QE	Using QE	Without QE	Using QE
Acc	19.89 %	**21.90 %**	13.64%	**18.99%**
MRR	9.12	**10.08**	6.06	**8.61**
AQ	27.45 %	**29.90%**	15.50%	**24.48%**

The results presented in Table XII above show that the Accuracy and the MRR have been improved for both the CLEF and TREC questions compared to what we have reached in the keyword-based evaluation before applying any relaxation at the answer checking stage. However, the number of answered questions has decreased. The reached accuracy and MRR are higher when we use our semantic QE together with JIRS.

The application of the different relaxations has significantly enhanced the different measures. The only exception is the decrease of the number of answered TREC questions when using JIRS without QE. The best performances are reached also with the use of JIRS on top of our semantic QE.

V. DISCUSSION

The experiments we conducted have shown that, regardless of the question set, the performance in terms of accuracy, MRR, and the number of answered questions improves when we include separately our semantic QE based on AWN and then JIRS as a structure-based PR system.

The highest performances are obtained when we include JIRS together with the semantic QE. Indeed, for the TREC questions the accuracy passes from 8.16% to 18.99%, the MRR from 3.1 to 8.61 and the percentage of the answered questions from 16.78% to 24.48% with respect to the different relaxations included at the answer checking stage.

The usefulness of using JIRS together with the semantic QE is better in the case of the CLEF questions. Indeed, the accuracy is close to 22% instead of 12%, the MRR reaches 10.08 rather than 3.85. The use of JIRS and the semantic

QE allows getting the answer in one of the first five returned passages for about 30% of the questions instead of 25.16%.

In order to analyze deeply the obtained performances, we have identified the number of the answered questions per type of question. Table XIV shows the result of this analysis for the two question sets.

TABLE XIV
TYPES OF THE ANSWERED QUESTIONS PER QUESTION SET (AFTER RELAXATIONS)

AQ				
	CLEF		TREC	
TYPES	Without JIRS+QE	Using JIRS+QE	Without JIRS+QE	Using JIRS+QE
ABREVIATION	6.49%	1.64%	2.78%	5.24%
COUNT	7.14%	8.74%	9.03%	5.71%
LIST	2.60%	2.73%	0.69%	0.95%
LOCATION	19.48%	21.86%	21.53%	22.38%
MEASURE	2.60%	1.64%	7.64%	5.24%
OBJECT	2.60%	2.19%	2.78%	4.76%
ORGANIZATION	5.19%	9.29%	6.94%	7.14%
OTHER	13.64%	12.57%	17.36%	13.33%
PERSON	29.87%	25.68%	14.58%	23.81%
TIME	10.39%	13.66%	16.67%	11.43%

The table presented above allows us to identify the question types that form the subset of the answered questions. For instance, for the CLEF questions, 25.68% of the answered questions are of the type PERSON and 21.86% are of the type LOCATION. For the CLEF questions, 80.87% of the answered questions are factoid ones while this percentage is 75.71% for the TREC set.

The difference of performances in term of accuracy, MRR and the answered questions between the two question sets can be explained by the fact that the TREC set contains a higher percentage of questions which does not belong to NE types (for instance LIST and OTHER).

Let us now consider the merged question set. Table XV shows the overall performances.

TABLE XV
THE OVERALL PERFORMANCES BEFORE AND AFTER USING THE SEMANTIC QE WITH JIRS (AFTER RELAXATIONS)

1,470 CLEF+TREC question		
MEASURES	Without JIRS and QE	Using JIRS+QE
Acc	9.66%	**20.20%**
MRR	3.41	**9.22**
AQ	20.27%	**26.74%**

The table above shows the usefulness of our two fold approach for the improvement of Arabic PR with respect to considered questions. Note that the results shown in this Table concern the 1,470 questions which can be expanded using our QE process so that we can compare performances before and after QE. The MRR has been enhanced more than the other measures. This means that the use of our approach increases the probability of having the expected answer in the first five ranked passages.

We have used the Student's paired t-test in order to show the significance of the obtained results. Therefore, for each measure (accuracy, MRR and the number of answered questions), we have considered the directional hypothesis that performances are better when we use JIRS together with our semantic QE approach. The null hypothesis is :

H_0 = There is no difference in performance (acc, MRR or #answered questions) either we use JIRS and QE or not.

Our alternative unilateral hypothesis is:
H_1 = The performance (acc, MRR or #answered questions) is better when we use JIRS and QE.

The t-test value is calculated using the following formula:

$$t = \frac{\left|\bar{x}_{no} - \bar{x}_{jq}\right|}{\sqrt{\frac{s^2(no)}{n(no)} + \frac{s^2(jq)}{n(jq)}}} \quad (8)$$

Where:
$\bar{x}_{no}$ is the mean (in terms of the considered measure) of the sample processed without using QE and JIRS.
$\bar{x}_{jq}$ is the mean (in terms of the considered measure) of the sample processed using QE and JIRS.

s^2 is the variance of the sample

n(S) is the number of observations in the sample S. In our case, we take into consideration four observations related to the different question collections used previously (858 TREC questions, 612 CLEF questions, 82 TREC questions and 82 CLEF questions).

The degree of freedom is: df = 7

The calculated t-test values are:

- In the case of accuracy: t=3.42
- In the case of MRR: t=1.45
- In the case of the Number of answered questions: t=2.23

According to the t-test values above, we can reject the null hypothesis in the case of the accuracy (t=3.42, df=7, $p<0.05$) and the number of answered questions (t=2.23,

df=7, $p<0.05$). For the MRR, the difference in performances between the two samples is not significant.

Therefore, our approach of combining semantic QE (using AWN) and JIRS at the PR stage is proofed to improve both the accuracy and the number of answered questions.

The results we obtained are encouraging in light of the following limitations:

- The low coverage of AWN which is the semantic resource used in our QE process;
- The experiments are conducted in an open domain (the web);
- The snippets returned by the Yahoo API are so small that it is difficult to have both the question terms and the expected answer in the same snippets;
- Questions do not come from the Arabic culture. Indeed, the CLEF and TREC questions used in the test are translated from the European and the American cultures respectively to the Arabic language. Hence, we are not sure that the available Arabic content in the web will cover or not the questions topics. This will cause a low redundancy level. Unfortunately, JIRS works better when redundancy is high, because it is more likely to retrieve at least one relevant passage in this case.
- Most of the answers are NEs that are transliterated from English or French to the Arabic language. Therefore, answers could not be found in Arabic texts and the performances can be affected by spelling errors.
- Generally, SE ranks the snippets according to their date of publication. Therefore, since our question sets belong to at most the year 2004, the answers could not appear in the first 1,000 considered snippets.

VI. Conclusion

In this paper, we have proposed an approach for enhancing passage retrieval for Q/A in Arabic. This approach tries to introduce the semantic aspect which is not included in the few existing Arabic Q/A projects. Indeed, the first step that we perform concerns the semantic QE based on the Arabic WordNet ontology. This step aims at reaching a high level of completeness by retrieving not only passages containing the question keywords but also those containing terms which are semantically related to them. Therefore, four semantic relations are considered: synonymy, hypernymy, hyponymy and the SUMO concept definition. The second step of our approach re-ranks the resulting passages with respect to their similarity to the question in terms of structure. The Distance Density n-gram model has been used at this step because its added value has been proved for other languages such as English. The JIRS PR system implements this model and is adapted to the Arabic language. Thus, the current work which uses JIRS has also the aim to confirm performances of JIRS in the context of the Arabic language.

The experiments that we have conducted using almost all available TREC and CLEF questions (2,264 questions) showed that the performances in terms of accuracy, the MRR and the number of answered questions has been improved significantly thanks to the use of our two steps approach. Indeed, the overall accuracy reached with the semantic QE used together with JIRS is 20.20% (versus 9.66%), the MRR passed from 3.41 to 9.22 and the number of answered questions from 20.27% to 26.74%. These results are encouraging since the experiments have been undertaken on the basis of the current release of AWN which has a low coverage of the considered questions (only 64.93% of the questions can be expanded). Moreover, the average of generated queries per question using the relation of synonymy (the most developed in AWN) is close to 4 queries. This average does not allow reaching a high level of completeness.

Experimenting with a more enriched release of the AWN ontology is among the intended future work. This would allow improving even more the performances of the passage retrieval on the basis of our approach.

Acknowledgment

We would like to thank Dr Yassine Benajiba for his remarks that have enriched this work.

References

[1] A. Ittycheriah, M. Franz, W.-J. Zhu, A. Ratnaparkhi, “IBM’s Statistical Question Answering System”. In Proceedings of the Ninth Text Retrieval Conference (TREC-2002). pp. 229-234.

[2] Abdelali, A., Cowie, J., & Soliman, H., 2006. “Improving query expansion precision using latent semantic analysis: Application on Arabic retrieval”. Journies d’Etudes sur le Traitement Automatique de la Langue Arabe (JETALA), Rabat, Morocco.

[3] Abouenour L., Bouzoubaa K., Rosso P. (2008) “Système de Question/Réponse dans le cadre d'une plateforme intégrée: cas de l'Arabe”. (in French) In: Proc. Rencontre Nationale en Informatique : Outils et Applications, RINOA-2008, Errachidia, Morocco, June 5-7.

[4] Abouenour L., Bouzoubaa K., Rosso P. “Improving Q/A using Arabic WordNet”. In: Proc. Int. Arab Conf. on Information Technology, ACIT-2008, Hammamet, Tunisia, December, 16-18.

[5] Abouenour L., Bouzoubaa K., Rosso P. “Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system” . In: Proc. Workshop on Computational Approaches to Semitic Languages, E-ACL-2009, Athens, Greece.

[6] Abouenour L., Bouzoubaa K., Rosso P., 2009. “Three-level approach for Passage Retrieval in Arabic Question /Answering Systems”. In Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco, May, 2009.

[7] Benajiba Y., Mona D., Rosso P. Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition. In: IEEE Transactions on Audio, Speech and Language Processing. Special Issue on Processing Morphologically Rich Languages, Vol. 17, No. 5, July 2009.

[8] 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.

[9] Benajiba Y., Rosso P., Lyhyaoui A., 2007. "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.

[10] Brini W., Ellouze M., Hadrich Belguith L. (2009). "QASAL : Un système de question-réponse dédié pour les questions factuelles en langue Arabe". In: 9ème Journées Scientifiques des Jeunes Chercheurs en Génie Electrique et Informatique, Tunisia. (in French).

[11] Buckwalter, T. (2004). "Issues in Arabic Orthography and Morphology Analysis". In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva.

[12] Buscaldi D., Rosso P., Gómez J.M., Sanchis E. "Answering Questions with an n-gram based Passage Retrieval Engine". In: Journal of Intelligent Information Systems (82) (in press) DOI: 10.1007/s10844-009-0082-y.

[13] Buscaldi, D., Gómez, J. M., Rosso, P., Sanchis, E., 2006. "The UPV at QA@CLEF 2006". In Working Notes for the CLEF 2006 Workshop.

[14] C. Amaral, H. Figueira, A. Martins, P. Mendes, C. Pinto, "Priberams Question Answering System for Poteguese". In Working Notes for the CLEF 2005 Workshop. In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Paper. Vol. 4022 of Lecture Notes in Computer Science, pp. 410-419. Springer, 2006.

[15] C. Monz. "From Document Retrieval to Question Answering". Ph.D. dissertation, Institute for Logic, Language, and Computation, University of Amsterdam, 2003.

[16] 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.

[17] F. Llopis, J. L. Vicedo, A. Ferrandez, "Passage Selection to Improve Question Answering". In Proceedings of the COLING 2002 Workshop on Multilingual Summarization and Question Answering, 2002.

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

[19] G. G. Lee, J. Seo, S. Lee, H. Jung, B.-H. Cho, C. Lee, B.-K. Kwak, J. Cha, D. Kim, J. An, H. Kim, K. Kim, 2002. "SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP". In Proceedings of the Tenth Text Retrieval Conference (TREC-2002). pp. 422-451.

[20] Gómez J. M., Buscaldi D., Rosso P., Sanchis E. "JIRS Language-independent Passage Retrieval system: A comparative study". In: Proc. 5th Int. Conf. on Natural Language Processing, ICON-2007, Hyderabad, India, January 4-6.

[21] 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.

[22] Gómez J., Buscaldi D., Bisbal E., Rosso P., Sanchís E. 2006. "QUASAR: The Question Answering system of the Universidad Politécnica de Valencia". In: Accessing Multilingual Information Repositories, Revised Selected Papers CLEF-2005, Springer-Verlag, LNCS(4022), pp. 439-448.

[23] Hammo, B., Abuleil, S., Lytinen, S., & Evens, M., 2004. "Experimenting with a question answering system for the Arabic language". Computers and the Humanities, 38(4), 379–415.

[24] 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.

[25] João Pinto F., "Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model". Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos, Vol. 2, No. 2, 2008.

[26] Kaszkiel, M. and Zobel, J., 2001. "Effective ranking with arbitrary passages". Journal of the American Society forInformation Science and Technology, 52(4):344-364.

[27] Larkey, L.S., Ballesteros, L., Connell, M.E., "Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis", In Proceedings of ACM SIGIR, pp. 269-274, (2002).

[28] Lauser B., "From thesauri to Ontologies: A short case study in the food safety area in how ontologies are more powerful than thesauri From thesauri to RDFS to OWL". Agricultural Information and Knowledge Management Papers, 2004.

[29] Mark A. Greenwood., 2004. "Using pertainyms to improve passage retrieval for questions requesting information about a location". In the ACM Special Interest Group on Information Retrieval (SIGIR) conference. Sheffield, UK, July 29th, 2004.

[30] Matthew W. Bilotti, Boris Katz and Jimmy Lin. "What Works Better for Question Answering: Stemming or Morphological Query Expansion?". In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004. 2004.

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

[32] Nanba H. 2007. "Query Expansion using an Automatically Constructed Thesaurus". In Proceedings of NTCIR-6 Workshop Meeting, May 15-18, 2007, Tokyo, Japan.

[33] 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.

[34] Risuh Ahn, Beatrix Alex, Johan Bos, Tiphaine Dalmas, Jochen L. Leidner, and Matthew B. Smillie. 2004. "Cross-lingual question answering with qed". In Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004), Bath, UK.

[35] Rodríguez H., Farwell D., Farreres J., Bertran M., Alkhalifa M., Antonia Martí M., Black W., Elkateb S., Kirk J., Pease A., Vossen P., and Fellbaum 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.

[36] Salton, G., Allan, J., Buckley, C., 1993. "Approaches to passage retrieval in full text information systems". In R. Korfhage, E. Rasmussen, & P. Willet (Eds.), Proceedings of the 16th annual international the ACM Special Interest Group on Information Retrieval (SIGIR) conference on research and development in information retrieval, Pittsburgh, PA (pp.49-58), New York: ACM.

[37] Seher I., 2006. "Query Expansion in personal queries". In IADIS Virtual Multi Conference on Computer Science and Information Systems (MCCSIS 2006) 15 – 19 May 2006.

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

[39] X. Liu and W. Croft., 2002. "Passage retrieval based on language models". In Proceedings of the Eleventh Internationa Conference on Information and Knowledge Management.

[40] Yonggang Qiu , Hans-Peter Frei, "Concept based query expansion", Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, p.160-169, June 27-July 01, 1993, Pittsburgh, Pennsylvania, United States.