Construction of an ontology for intelligent Arabic QA systems leveraging the Conceptual Graphs representation

Lahsen Abouenour[a,*], Mohamed Nasri[a], Karim Bouzoubaa[a], Adil Kabbaj[b], Paolo Rosso[c]

[a] *Computer Science Department, Mohammadia School of Engineers, Mohammed V-Agdal University, Rabat, Morocco*

[b] *Computer Science Department, INSEA, Rabat, Morocco*

[c] *NLE Lab, PRHLT Research Center, Universitat Politècnica de València, Spain*

Abstract. The last decade had known a great interest in Arabic Natural Language Processing (NLP) applications. This interest is due to the prominent importance of this 6th most wide-spread language in the world with more than 350 million native speakers. Currently, some basic Arabic language challenges related to the high inflection and derivation, Part-of-Speech (PoS) tagging, and diacritical ambiguity of Arabic text are practically tamed to a great extent. However, the development of high level and intelligent applications such as Question Answering (QA) systems is still obstructed by the lacks in terms of ontologies and other semantic resources. In this paper, we present the construction of a new Arabic ontology leveraging the contents of Arabic WordNet (AWN) and Arabic VerbNet (AVN). This new resource presents the advantage to combine the high lexical coverage and semantic relations between words existing in AWN together with the formal representation of syntactic and semantic frames corresponding to verbs in AVN. The Conceptual Graphs representation was adopted in the framework of a multi-layer platform dedicated to the development of intelligent and multi-agents systems. The built ontology is used to represent key concepts in questions and documents for further semantic comparison. Experiments conducted in the context of the QA task show a promising coverage with respect to the processed questions and passages. The obtained results also highlight an improvement in the performance of Arabic QA regarding the c@1 measure.

Keywords: Ontology, Conceptual Graphs, Arabic Natural Language Processing, Question Answering, WordNet, VerbNet

1. Introduction

Recently, many surveys highlighted the impressive growth of the Arabic content on the Web that currently plays a key role in people's life and companies' strategies. This reflects the increasing number of users interesting in all kinds of information expressed in Arabic. The processing of such great amount of content is necessary to reduce information overload.

Nowadays, computers present notable possibilities in terms of storage and time processing capabilities. Nevertheless, they are still less effective when it comes to understand the meaning of written languages.

Ontologies are among the resources that can allow computers for understanding the meaning of texts and, in turn, leveraging their capabilities to develop more sophisticated applications for end users. This kind of resources is used in various fields including Natural Language Processing (NLP), information retrieval, machine learning, data mining, and knowledge representation. Basically, Gruber [14] defines ontologies as formal and explicit specifications in the form of concepts and relations of shared conceptualizations.

Currently, users exploit the great amount of available information on the Web through Search Engines (SEs). Nevertheless, such SEs are not suitable for advanced users' needs, especially when they

[*] Corresponding author. E-mail: abouenour@yahoo.fr

look for answering questions rather than getting lists of documents about a topic. Research in the NLP field is concerned with providing systems satisfying these specific needs.

The last decade was particularly active in terms of number and quality of Arabic NLP research. Consequently, some basic challenges related to the processing of this language have significantly been resolved. Among these challenges, we can cite high inflection and derivation, Part-of-Speech (PoS) tagging, and diacritical ambiguity for which maturity was respectively reported in [29, 9, 15, 7].

On the other hand, the development of more sophisticated systems has witnessed just a few attempts and low levels of maturity in comparison with existing systems for other languages. This is for example the case of Question Answering (QA) systems that theoretically go beyond the classical retrieval of lists of documents and try to automatically answer natural language questions. The main advantage of QA systems is their ability to reduce the complexity faced by users when looking for such answers on the Web or within other collections of documents [34-37]. In fact, users of QA systems are not obliged to view dozens of documents related to returned snippets to obtain the expected answer.

To build an efficient QA system, there are many requirements in terms of integration of different NLP tasks such as Query Expansion (QE), word sense disambiguation, Named Entity Recognition, etc. This integration has the aim to obtain a high level of semantic understanding capabilities. Indeed, the system would need to recognize the features and morphology of each question term as well as relations between these terms and the syntactical and semantic representation of the question. To perform such process, an ontology with a high coverage of lexical terms as well as semantic representation of concepts is needed. Indeed, ontologies can support deeper approaches based on semantic understanding as reported in many QA research works [13, 2].

However, there is currently no Arabic ontology with the mentioned features. The only similar resource for this language is the Arabic WordNet (AWN)[1] [10] that represents semantic relations between synsets (groups of synonyms). The latters are not assigned formal definitions as done in ontologies.

In this paper, we present a new built ontology for NLP applications, especially Arabic QA systems. The main objective behind this work is making available an ontology allowing semantic representation of key concept meanings to support semantic reasoning and intelligent systems. This paper shows how these frames have been transformed into the CG formalism for a better semantic representation and matching in intelligent Arabic QA systems.

The targeted features for this ontology are: (i) high coverage of the Modern Standard Arabic (MSA) to process texts written in MSA, and (ii) large semantic and hierarchical relations between concepts to perform efficient QE and semantic-based processing. Therefore, existing resources with such high coverage of the MSA and relation features are used. Ontology entries were acquired from synsets and semantic relations in the AWN lexical resource. The syntactical and semantic frames of Arabic VerbNet (AVN)[2] [23, 24] were the basis of verb meaning representation in the built ontology.

The remainder of this paper is structured as follows. Section 2 describes the main works related to ontology construction, highlighting Arabic resources especially AWN and AVN. Section 3 is devoted to the description of the three steps of the ontology building process. Section 4 presents the experiments conducted and the evaluation made to validate the usefulness of this new ontology in the Arabic QA task. Section 5 draws the main conclusions of this work and lists further works.

2. Related works

Beyond the importance of ontologies in semantic reasoning-based and intelligent QA systems, another key element in such systems is the formalism used to represent knowledge in questions and documents. In this case, the inspiration comes from existing approaches where Conceptual Graphs (CGs) [31] have been used. A CG is a directed graph of nodes that correspond to concepts, connected by labeled and oriented arcs that represent conceptual relations. The CG formalism has the advantage to be close to natural language and can be manipulated by computers.

Among the approaches having adopted CGs, we can mention those based on the semantic similarity scoring. Montes-y-Gómez et al. [22] proposed a technique for this CG-based similarity. Hensman and Dunnion [16] showed its usefulness to automatically represent questions and passages and to improve precision in QA. Their approach used VerbNet (VN) [19] and WordNet (WN) [12].

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

[2] http://ling.uni-konstanz.de/pages/home/mousser/files/Arabic_verbnet.php

For languages such as English, there are many interesting open source ontologies that belong either to the specific or open domain category: OpenCyc [21], Know-ItAll [11], HowNet[3], SNOMED[4], GeneOntology[5], etc. Also, there are many ontologies that are dedicated to a specific type of information. This is the case of the YAGO ontology with around 3,000,000 Named Entities (NEs) and 20 millions of facts about them.

There are also many top level ontologies that are language-independent since they contain just the main common concepts shared across languages. The SUMO ontology [28] is an example of such general ontology. An extension to the SUMO ontology considered the integration of some mid-level concepts devoted to the Arabic culture. Unfortunately, such attempts in building Arabic-oriented ontologies present the lack of containing just a few number of concepts.

The AWN resource was an alternative to this lack by containing larger number of entries. It was constructed on the basis of the methods developed for Princeton WordNet (PWN) [12] and EuroWordNet [32]. The AWN project succeeded to come up with a linguistic and semantic resource that complies with the WN structure while considering some specificities of Arabic such as diacritized entries. This resource is freely available and as such fill in the gap in the Arabic NLP community. The structure of AWN is similar to an ontology since its entries are connected through semantic relations particularly the hypernymy/hyponymy relation.

Although AWN contains lexical information about synsets (PoS, words, roots, etc.), their meaning is not represented in this resource except the corresponding English gloss. Unlike AWN, the recently developed VerbNet for Arabic (i.e., AVN) fills in this gap [23, 24]. This new resource has the particularity to provide a classification of Arabic verbs using Levin's classes [20], integrate frames about verb syntax and semantics and give some lexical information about these verbs.

Hence, a combination between the two complementary aspects of AWN and AVN would be helpful for the Arabic NLP community. This idea has been already investigated in other languages. In fact, Pazienza et al. [25] created a resource by mixing sense relational knowledge enclosed in English WN, frame knowledge enclosed in VN and corpus knowledge enclosed in PropBank [8]. The created resource helped then in increasing the F-measure up to 85% in the context of Textual Entailment acquisition for QA. In a similar direction, Kaisser [26] showed the usefulness of combining the three previously mentioned resources to correctly answer over 62% of the 500 considered questions from TREC-2002[6].

3. Ontology construction process

To our knowledge, there has been no combination of lexical and semantic resources for the Arabic NLP. The main objective of the current research is to propose a new ontology built from AWN and AVN that can be used in semantic-based processing such as Arabic QA systems.

The design of our ontology is structured around a concept hierarchy, lexical information and situations about these concepts. Figure 1 illustrates this design where the proposed ontology considers two main kinds of information: (i) Concepts hierarchy and lexical information, and (ii) Situations related to concepts. The former are extracted from AWN, whereas the latter are represented based on the transformation of syntactic and semantic AVN frames into CGs. This transformation will be detailed in a further section.

In Figure 1, boxes with bold lines refer to concepts and their hierarchy, boxes with gray background refer to additional information about concepts and boxes with dashed lines refer to situations. The ontology root is the most general concept of the ontology. Under this general concept we can find the other concepts extracted from AWN synsets. Each concept can have hyponymy relation with other concepts that are more specialized (their meanings are more specific). The lexicon is the natural language counterpart of the concept, i.e., the words referring to this concept in the considered language (Arabic in our case). For example the concept “رأى” (to see) can be expressed in the Arabic lexicon by one of the following words: نظر, رأى, حدق, etc. The concept itself has another sub concept which is a specialization of “رأى”, namely “راقب” expressed in natural language by: راقب, نظر بتركيز, جال بنظره, etc. According to the different expressions, in natural language, of the same concept, we can have syntactic-semantic situations that can be applied to this concept: for example, a situation where the syntax contains

[3] www.keenage.com/html/e_index.html
[4] www.snomed.org
[5] www.geneontology.org

[6] http://trec.nist.gov

V+Agent+Patient (for instance, راقب النظار الهلال) with a specific meaning, another where the syntax contains V+Agent+Patient+PP (for instance, رأى الشرطي اللص في الليل), etc. The situations are simply use cases of the concept from two perspectives: syntax and semantic. Each situation refers to a syntax case together with the corresponding semantic meaning. These situations are translated into CGs as described in the following section.

Note that the ontology contains not only static information (concepts, lexicon and situations) but also dynamic information, for example NEs, i.e., instances (or individuals) such as names of persons and places that are important to be recognized by real applications including QA systems. Indeed, this is important in the case of factoid questions where the expected answer is a NE. Therefore, we extended in a previous work the original AWN by mapping a large number of NE from the YAGO ontology with AWN synsets [3].

In the following sub sections, we highlight the process performed to populate our ontology with respect to the above two kinds of information.

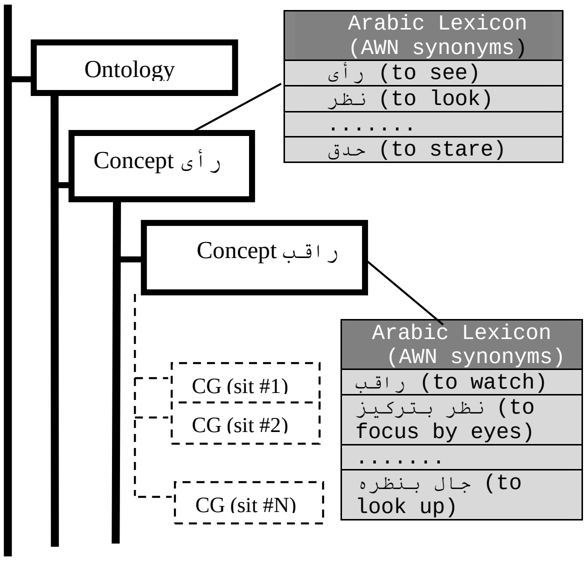


Fig. 1. Design of the proposed ontologyConcept hierarchy and lexical information

The hierarchy of our ontology, built from AWN synsets and the hypernymy relations between these synsets, is composed of two parts:

- The first one contains concepts related to AWN synsets. Actually, each synset was transformed into a concept. Two nodes are then created: "verb" and "noun" nodes. These concepts are also assigned a lexicon representing the words that can express the given concept in real world texts. This lexicon is language-dependent and is provided by AWN words that are members in the given synset. The integrated lexicon considers both the voweled and unvoweled forms of words. For example, as illustrated in Figure 1 the entry related to "Concept 1" is associated with the lexicon composed of the words: "رأى" (unvoweled form of raOaY), "نظر" (nZr, synonym of raOaY) and "حدق" (Hdq, synonym of raOaY). In that example, "Concept 2" is a sub type of "Concept 1" which means that the former is more specialized than the latter (more general concept). In addition to this lexicon, concepts under the node "verb" are also assigned situations that are transformations of the syntactic and semantic AVN frames into CGs. These transformations are detailed in the next section.
- The second part of the ontology hierarchy contains nodes regrouping different additional concepts and relational concepts that are needed to express the situation CGs. For example, the node "linguistic" contains the concepts "verb", "noun_phrase", "preposition", etc; the node "action_root" contains AVN semantic predicates; etc.

In the next section, we provide details about the usage of the latter part to build CG situations from AVN frames and how these are inserted in the corresponding concept nodes of the former part.

3.1. *Ontology concept situations*

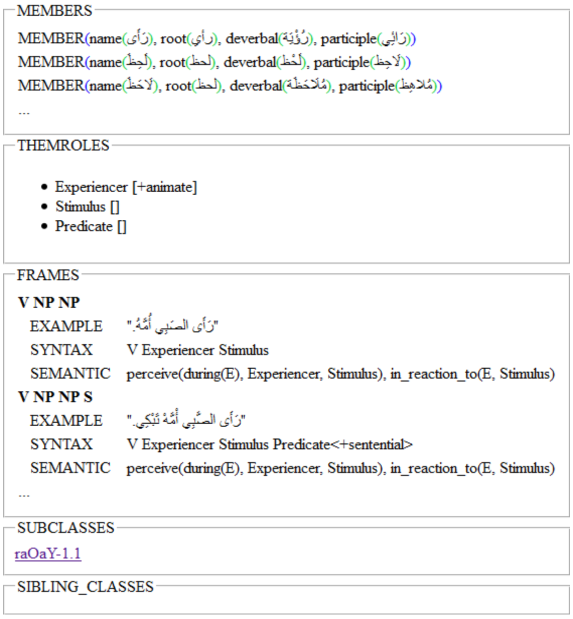
3.1.1. *Arabic VerbNet structure*

The concepts related to the "verb" node are assigned CG situations corresponding to AVN semantic and syntactical frames. Let us recall that AVN is a large coverage verb lexicon exploiting Levin's classes [20] and the basic development procedure of Kipper Schuler [19]. The current version has 336 classes populating 7744 verbs and 1399 frames[7]. Figure 2 shows the AVN content related to class raOaY-1 (i.e., رأى).

As depicted in Figure 2, every class contains information about: (i) class members (i.e., verbs belonging to the class), for instance رأى (to see), لاحظ (to observe), etc. (ii) themeroles and frames that represent syntactic-semantic situations of its members (for example, V Experiencer Stimulus),

[7] http://ling.uni-konstanz.de/pages/home/mousser/files/Arabic_verbnet.php

and eventually (iii) its sub classes and sibling classes (in the above example, the sub class is raOaY-1.1 and there is no Sibling class).



MEMBERS

MEMBER(name(رَأَى), root(رأي), deverbal(رُؤْيَة), participle(رَائِي))
MEMBER(name(لَحِظَ), root(لحظ), deverbal(لَحْظ), participle(لَاحِظ))
MEMBER(name(لَاحَظَ), root(لحظ), deverbal(مُلَاحَظَة), participle(مُلَاحِظ))
...

THEMROLES

- Experiencer [+animate]
- Stimulus []
- Predicate []

FRAMES

V NP NP
EXAMPLE "رَأَى الصَّبِي أُمَّهُ."
SYNTAX V Experiencer Stimulus
SEMANTIC perceive(during(E), Experiencer, Stimulus), in_reaction_to(E, Stimulus)
V NP NP S
EXAMPLE "رَأَى الصَّبِي أُمَّهُ تَبْكِي."
SYNTAX V Experiencer Stimulus Predicate<+sentential>
SEMANTIC perceive(during(E), Experiencer, Stimulus), in_reaction_to(E, Stimulus)
...

SUBCLASSES

raOaY-1.1

SIBLING_CLASSES

Fig. 2. A snapshot of the AVN class raOaY-1

In more details, the top level of each class shows the verbs that are members of the given class. Each verb member is identified by the verb itself, its root form, its deverbal form and its participle. Also, the thematic roles and their restrictions are encoded at the top level of classes; restrictions are lists of selectional constraints on semantic roles. Some frames define local restrictions that are specific to the given frame and are combined with the common restrictions (i.e., those appearing at the top level of a class).

Frames related to a given class are presented with an example sentence, a syntactic and a semantic structure. The latter structure contains semantic predicates including arguments and temporal information similarly to that proposed by Moens and Steedman [27].

Sub classes (for instance raOaY-1.1) have a similar structure as the main classes (i.e., raOaY-1). Obviously, sub classes can also have sub classes in a recursive way. A sub class inherits all properties of the main class. Therefore, verbs appearing in these sub classes have new syntactic and semantic frames in addition to those of the main class. On the other hand, sibling classes are specific to the Arabic language and are detailed in the work proposed by Mousser [23].

3.1.2. Transformation of AVN frames into CGs

The structure and content of AVN classes is an interesting starting point to enrich the verb nodes of our ontology using semantic and syntactic information. To achieve this enrichment, we perform a two-step approach.

As shown in Figure 3, the two following steps are performed:

- Step 1: The first step is concerned by the extraction, from AVN, of verbs together with corresponding frames content. A given verb can appear as member of different classes. Therefore, we extract the frames from all these classes as well as from their super classes (considering the principle of frame inheritance).
- Step 2: we generate CGs based on the extracted semantic information and integrate them in the ontology as situations of each concept (corresponding to the concerned verb members). Figure 4 provides the general design of these CGs.

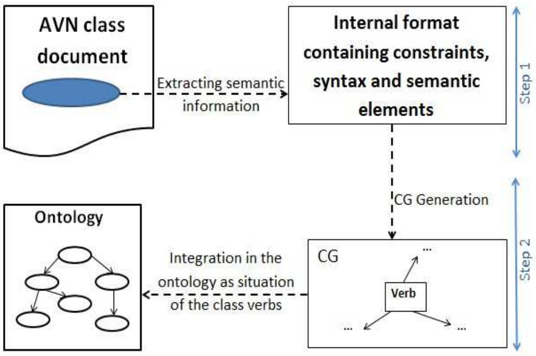

AVN class document
Extracting semantic information
Internal format containing constraints, syntax and semantic elements
Step 1
CG Generation
Ontology
Integration in the ontology as situation of the class verbs
CG
Verb
...
...
...
Step 2

Fig. 3. General architecture for the semantic extraction

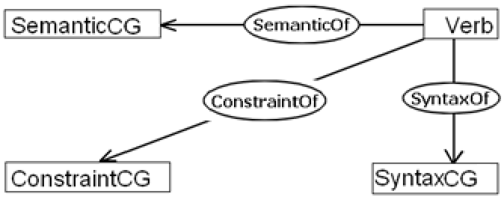


Fig. 4. Form of the situation CG corresponding to the AVN frame

Figure 4 depicts the three main sub CGs of the global CG that formalizes AVN frames: (i) “SyntaxCG” for the syntactic frame that can be applied to a given verb, (ii) “SemanticCG” for the meaning of the verb by means of themeroles and predicates, and (iii) “ConstraintCG” for the constraints existing on themeroles used in the first and second sub CGs. The global CG is represented as a verb concept linked to the other sub CGs through three ontology relations, respectively “SyntaxOf”, “SemanticOf” and “ConstraintOf”. Figure 5 details the illustration of the process performed to transform a frame component (i.e., syntactic frame, semantic frame and constraints) into sub CGs.

As we can see from Figure 5, the step “CG generation” (step 1 in Figure 3) is performed through the following five sub steps:

- Step 2.1: A given verb from AVN is located in the AWN ontology. This means that corresponding concepts are identified. A verb can be associated with different possible concepts. To disambiguate these possibilities, we consider the concept with the ontology lexicon containing the highest number of verbs that are members for the same class of the given verb.
- Step 2.2: For each syntactic frame extracted in step1, the succession of syntactic constituents such as Noun Phrases (NP) and Prepositional Phrase (PP) are represented in the “SyntaxCG” using general concepts (for instance the concept “np**”** connected through the ontology relation “followedBy”). Examples of resulting Syntactic CGs are provided below:

```
Syntactic CG1:
      [np : *c2 ] -
      -followedBy->[np : *c3 ],
      <-followedBy-[verb : *c1 ]

Syntactic CG2:
      [np : *c2 ] -
           -followedBy->[np : *c3 ]-
            followedBy->[np: *c4],
           <-followedBy-[verb : *c1 ]
```

These CGs concern the two frames of the AVN class illustrated above in Figure 2.

- Step 2.3: We construct “ConstraintCG” from restrictions that are both common among the entire class as well as those that are specific to the given frame. The following CG is the “ConstraintCG” generated for the class illustrated in Figure 2 :

```
Constraint CG:
               [list : "[?c2(animate)]"]
```

As can be noticed, the CG representation only contains the restriction on the themerole Expertiencer and does not consider the restriction on the theme role Predicate [+sentential]. Indeed, this theme role is not used in the frames of the given class. The resulting CG shows that the constraint on the concept of type “np” and identified by “c2” in the syntactic CG2 is: it must be “animate”.

- Step 2.4: The CGs corresponding to the semantic frames are constructed by means of a semi-automatic process. Let us take the same AVN class illustrated in Figure 2. The first semantic frame shows that:
  - During the event related to verbs that are members of the given class, the syntactic constituent “Experiencer” (i.e., the second NP referenced by “c2” in SyntaticCG1) perceives the syntactic constituent “Stimulus” (i.e., the third NP referenced by “c3”);
  - This event is in reaction to the syntactic constituent “Stimulus”.

Hence, the two above constituents of the semantic frame are represented in the semantic CG as follows:

```
Semantic CG:
   [event : *p1 ]-
          -duringOf->[cg:[perceive:*p2 ]-
          -experiencerOf->[np : ?c2 ],
          -stimulusOf->[np : ?c3 ]],
          -inReactionTo->[np : ?c3 ]
```

In the above semantic CG, the references used in the syntactic and constraint CG are reused for the same constituents in order to make a connection between parts of the global CG (illustrated in Figure 4).

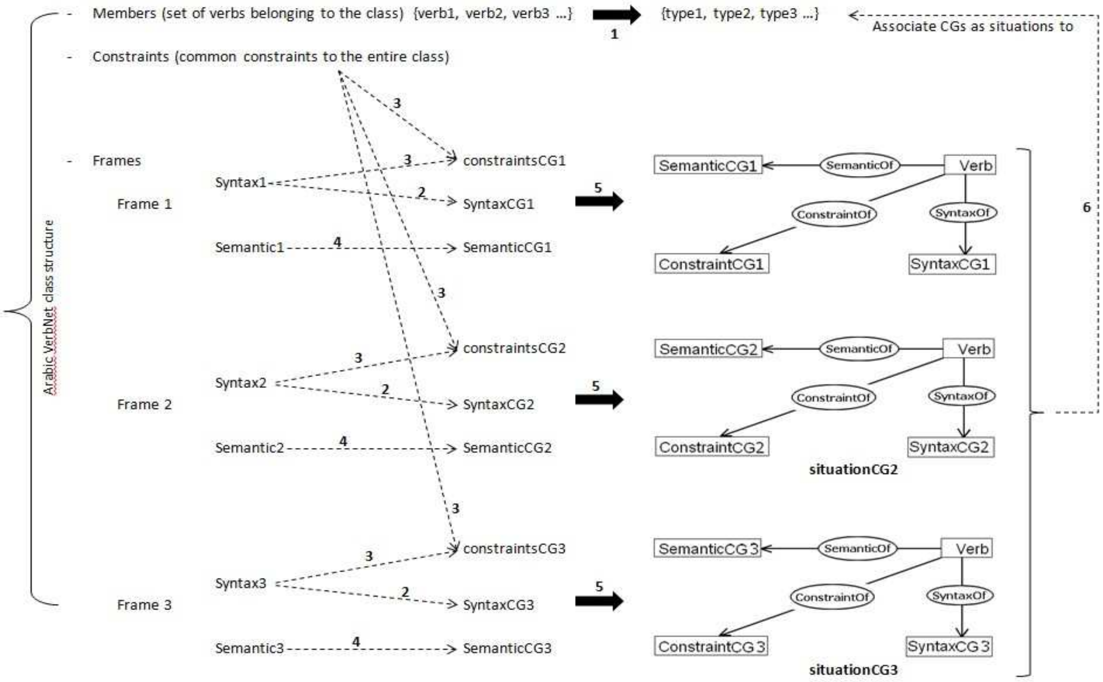


Fig. 5. . Different steps for CG generation

As shown in the semantic CG, the two AVN predicates "*perceive*" and "*in_reaction_to*" are represented differently: the former becomes the concept "*perceive*" whereas the latter becomes the relation "*in_reaction_to*". The decision of which representation form should be used (concept or relation) is made manually. Thereafter, many types of automatic transformation generate the resulting CG. This was applied to the 146 different predicates contained in AVN as shown in Table 1.

Table 1

Transformation of AVN predicates into semantic CGs

AVN predicate groups	Example	No. predicates	No. transformation types
group 1	adopt, allow, attempt, contact,	87	1
group 2	free, depend, meet	39	39
group 3	together-apart, harmed-disconfort	8	4
group 4	-	3	1

Table 1 shows that 87 (about 60%) of the available predicates are mapped using the same semi-automatic algorithm. The remaining ones are processed according to 3 groups: the first group contains 39 predicates (About 27%) that are mapped using 39 different algorithms (the manual task in this case is repeated 39 times); the second group only concerns 8 predicates with 4 different algorithms (one per predicate pair); finally, 3 other predicates required another algorithm.

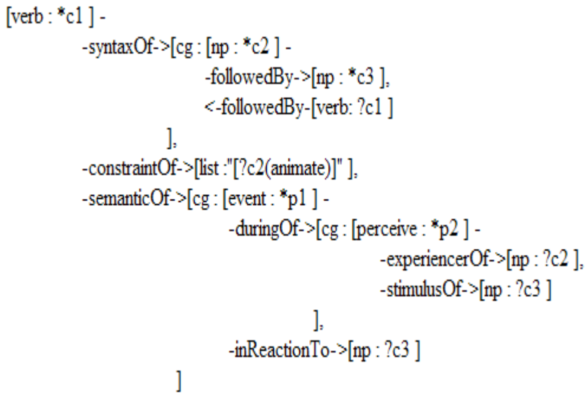
- Step 2.5: We construct the global CG as explained above (Figure 4).
- Step 2.6: The resulting global CG is associated with concept extracted after Step 1. The general concept "verb" is substituted in this global CG by each associated concept.

After transforming the AVN frames into CGs as described above, we use the Amine Platform [17] to implement the AWN-AVN ontology. The resulting CGs are stored under the situations of the corresponding ontology concept.

The construction of our ontology is made within this platform which is a Java open source multi-layer platform dedicated to the development of intelligent systems and multi-agents systems [18]. In addition to these characteristics, Amine has been chosen due to its features, namely: (i) its support of the CG formalism, and (ii) its modular environment providing an Ontology layer that we use for manipulating the AWN ontology, an Algebraic layer with various matching-based operations (like match, equal, unify, subsume, compare, maximalJoin, generalize, analogy, etc.) and a Knowledge Base (KB) support for advanced semantic processing.

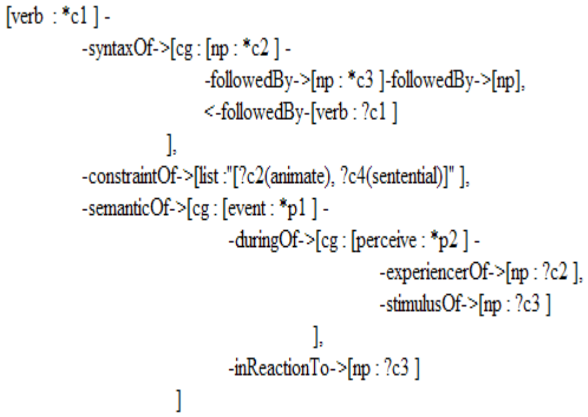
Here are the two CGs corresponding to the two frames of the previous example (class raOaY-1):

Global CG 1:



```
[verb : *c1 ] -
          -syntaxOf->[cg : [np : *c2 ] -
                          -followedBy->[np : *c3 ],
                          <-followedBy-[verb: ?c1 ]
                    ],
          -constraintOf->[list :"[?c2(animate)]" ],
          -semanticOf->[cg : [event : *p1 ] -
                              -duringOf->[cg : [perceive : *p2 ] -
                                                -experiencerOf->[np : ?c2 ],
                                                -stimulusOf->[np : ?c3 ]
                                         ],
                              -inReactionTo->[np : ?c3 ]
                     ]
```

Global CG 2:



```
[verb  : *c1 ] -
          -syntaxOf->[cg : [np : *c2 ] -
                          -followedBy->[np : *c3 ]-followedBy->[np],
                          <-followedBy-[verb : ?c1 ]
                    ],
          -constraintOf->[list :"[?c2(animate), ?c4(sentential)]" ],
          -semanticOf->[cg : [event : *p1 ] -
                              -duringOf->[cg : [perceive : *p2 ] -
                                                -experiencerOf->[np : ?c2 ],
                                                -stimulusOf->[np : ?c3 ]
                                         ],
                              -inReactionTo->[np : ?c3 ]
                     ]
```

4. Experiments in the context of Arabic QA

We evaluated the usefulness of the described ontology from an applied NLP task perspective. Concretely, the ontology is used in a three-level approach based on keyword, structure and semantic reasoning levels respectively. This approach was previously described in [1-6]. In those works, experiments focusing on the first two levels were presented. In the current section, we present and discuss the new experiment we conducted for the semantic-based level that integrates the proposed ontology.

4.1. Experimental process

The aim of this experiment is to measure the ability of our ontology to support a semantic reasoning process in the context of Arabic QA. Briefly, our approach is based on two steps: (i) Step 1 that has the aim to automatically represent a question and candidate passages in terms of CGs, and (ii) Step 2 that measures the semantic similarity between the CG of the question and that of the passages. Step 2 performs a “Generalization” operation between question CG and each passage CG.

Before presenting the results obtained in the present experiment, we provide details about Step 1. This step is composed of five sub steps:

- Step 1.1 “*Words analysis*”: it analyzes words of the text by means of Al Khalil Morphological Analyzer[8]; All the candidate analysis of a word are considered to be validated in further steps;
- Step 1.2 “*Cross resource matching*”: we try to match the different stems returned by Al Khalil Analyzer with their corresponding verbs in the Arabic VerbNet resource. This matching is made through the verb, the deverbal or the participle attributes of the AVN verbs.
- Step 1.3 “Sub *CGs retrieval*”: Once the AVN verbs are identified, we extract from our ontology all the CGs related to these verbs; note that these CGs are transformations of the semantic and syntactic frames into CGs as described in Section 3.2.
- Step 1.4 “*Sub CGs disambiguation*”: The present step has a two-fold goal: (i) disambiguating the previous candidate CGs, that were extracted after the achievement of Step 1.3, in order to obtain a short list of final Sub CGs, and (ii) enriching (or instantiating) the final CGs according to the processed text (question or passage). The present step starts by a syntactic parsing of the text using the Stanford Parser[9] for Arabic. The Typed Dependencies (TD) provided by this parser allow us to construct dependencies CGs “dep-CG” after applying 11 syntax-to-CG rules. The sub CGs are then disambiguated against each “dep-CG” according to the “Join” operation between CGs.
- Step 1.5 “*CG construction*”: the final CG of the text is then constructed using the “MaximalJoin”

[8] Can be downloaded from http://sourceforge.net/projects/alkhalil/

[9] http://nlp.stanford.edu/software/index.shtml

operation between the Sub CGs kept after Step 1.4.

4.2. Test-set

The 2013 QA4MRE question set [33] is composed of 4 topics, namely “Aids”, “Climate change” and “Music and Society” and “Alzheimer”. Each topic includes 4 reading tests. Each reading test consists of one single document, with at least 15 questions and a set of five choices per question. There are 44 auxiliary questions that are duplicates of the main questions, but without required inference. This allows testing the ability of systems to use the inference technique and its impact in the question processing.

For each question in the test-set, we perform the first two levels of our approach. From the set of the resulting passages we extract a sub set of 15 passages that are assigned the best surface similarity score (i.e., score measured based on keywords and structure and without semantic processing). Thereafter, we perform, either for the question and the considered passages, the step of constructing CGs.

4.3. Results

4.3.1 Word analysis

After performing the five sub steps of CG representation (words analysis, cross resource matching, sub CGs retrieval, sub CGs disambiguation, CG construction), we analyzed the results obtained in each step. Regarding the word analysis step based on Al Khalil Morphological Analyzer, we noticed that all the questions and corresponding passages were concerned by at least one analysis solution.

In total, there are 27,073 solutions provided by Alkhalil Analyzer for the 284 questions (this represents an average of 95 possible solutions per question). Among these solutions, 63.7% of them correspond to nouns and 33.7% to verbs. These solutions contain 873 distinct stems. For each question, we extract the best 15 passages according to the surface similarity score (for 35% of the questions we could not extract more than 8 passages, the average of the extracted passages per question is 10). The morphological analysis of the 2,734 extracted passages provided 600,399 possible solutions. The distribution of these solutions over PoS (64.6% are nouns and 32.9% are verbs) is quite similar to the one registered in the questions. The number of distinct stems in these solutions is 4,308.

4.3.2 Cross resource matching

The cross resource matching recognized 322 question stems in the Arabic VerbNet resource and 4,306 stems in the corresponding passages. The details of this cross matching step are presented in Table 2.

Around 43% of the recognized stems in the questions were matched using the verb-matching, 46% approx. using the deverbal-matching and only 11% using the participle-matching. As for passages, there is a number of 1,252 matched stems which is lower in percentage (29%) than that registered for questions (37%). Nevertheless, the distribution of this number over the different types of matching is quite similar (41% using verb-matching, 44% using deverbal-matching and 15% using participle-matching).

Table 2

Cross resource matching statistics – AVN matching

	Questions		Passages	
	Number	%	Number	%
Distinct stems	873	-	4,306	-
Matched	322	37%	1,252	29%
Verb-matching	139	43%	511	41%
Deverbal-matching	147	46%	547	44%
Participle-matching	36	11%	194	15%

The second part of the cross resource matching step consists of considering the AWN part of our ontology as shown in Table 3.

Table 3

Cross resource matching statistics – AWN matching using Standard and Enriched versions

	No Distinct Stems	Matched in AWN	%
Questions	873	568	65.06%
Passages	4,308	2,324	53.95%

The AWN synsets integrated in our ontology cover around 65% of question stems and roughly 54% for passages. This shows the effectiveness of the integrated AWN content for the application of the different steps of our semantic-based approach.

4.3.3 Syntactic parsing-based CGs

In this step, we performed the syntactic parsing by means of the Stanford parser for the set of 284 questions and their corresponding 2,734 passages. The parsing of the passages was preceded by splitting them into phrases in order to increase the accuracy of parsing. The statistics of the questions and passages that were matched by our typed dependencies rules are listed and illustrated below.

Table 4 shows the high coverage of the Stanford parser that allowed getting parsing solutions for around 98.6% of the questions and 83.1% of the passages. For the remaining questions and passages, the parser could not process the text due mainly to the limit reached in terms of text length despite the splitting of passages into phrases.

Table 4

Applied typed dependencies rules for questions and passages

	Questions (Q)		Passages (P)	
	Number	%	Number	%
Set	284	-	2,734	-
Matched	280	**98.59%**	2,272	**83.10%**
TDs	2,632	-	25,008	-
TDs matched	1,473	**55.97%**	15,156	60.60%
- Rule #1	69	4.68%	6,471	42.70%
- Rule #2	152	10.32%	14,229	93.88%
- Rule #3	222	15.07%	7,786	51.37%
- Rule #4	196	13.31%	7,734	51.03%
- Rule #5	255	**17.31%**	14,274	**94.18%**
- Rule #6	174	11.81%	7,215	47.60%
- Rule #7	9	0.61%	355	2.34%
- Rule #8	40	2.72%	1,837	12.12%
- Rule #9	72	4.89%	7,346	48.47%
- Rule #10	222	15.07%	12,357	81.53%
- Rule #11	62	4.21%	3,432	22.64%
Rules overlap rate	4.82%		8.76%	

For both questions and passages, all the 11 rules were applied at least once.

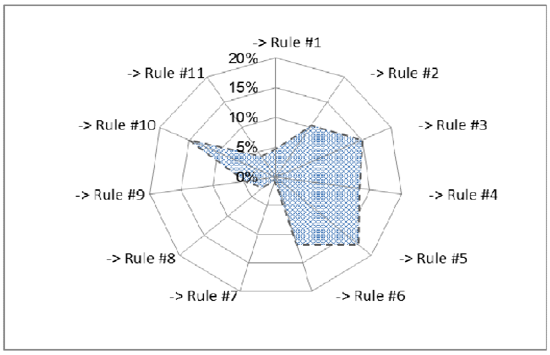


Fig. 6. Distribution of question' typed dependencies over rules

As illustrated in Figure 6 and Figure 7, the most applied rule is Rule #5 in both sets (17.31% of the matched TD in question and 94.18% in passages). The ranking of 4 rules (Rule #1, Rule #7, Rule #8 and Rule #11) is also the same in both sets. Note that for around 5% of question TD and 43% of passage TD, more than one rule was applied for the same TD. This is due to the fact that in some cases two rule conditions are matched in the given TD. This mainly concerns Rule 1# and Rule #9.

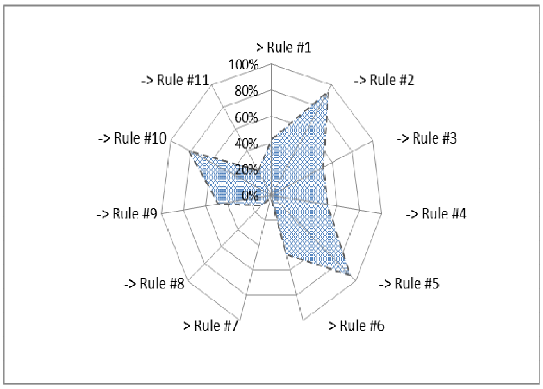


Fig. 7. Distribution of passages' typed dependencies over rules

4.3.4 Semantic similarity

After representing the question and the candidate passages in terms of CGs, we calculate the semantic similarity score proposed by Montes-y-Gomez [22] between both CGs. Thereafter, we measure the performance of the system (i) using a surface-based approach [1-6] and (ii) after using the semantic approach described in Section 4. Table 5 and Table 6 display the obtained results.

As we can see, the semantic-based approach that uses the CG representation (constructed through the built ontology) improves the performance in terms of the percentage of correctly answered questions from 7.39% out of 284 questions to 16.2%.

Table 5

System performance with surface-based approach

	Number	%	Remark
Questions	284	-	
Avg(surface similarity)	0.319		
Questions answered by IDRAAQ (based on AWN)	164	57.75%	out of 284
Questions unanswered by IDRAAQ (based on AWN)	120	42.25%	out of 284
Questions correctly answered	21	7.39%	out of 284
		12.80%	out of 164
C@1	0,11	-	-

Another aspect that deserves to be mentioned is the high percentage of questions that were given an answer by the system (77.11% versus 57.75%). The improvement was also registered regarding the C@1 measure that penalizes if a system provides wrong answers. The semantic-based approach obtained 0.20 in terms of the C@1 measure (versus 0.11 with the surface-based approach).

Table 6

System performance with semantic-based approach

	Number	%	Remark
Questions	284	-	
Avg(semantic similarity)	0.697	-	
Questions answered by IDRAAQ (based on AWN)	219	77.11%	out of 284
Questions unanswered by IDRAAQ (based on AWN)	65	22.89%	out of 284
Questions correctly answered	46	16.20%	out of 284
		21.00%	out of 219
C@1	0,20	-	-

We also note that the average similarity increased using the semantic similarity score (0.697 versus 0.319 for the surface similarity). This means that based on CG comparison it was possible to identify the similarity between the question and the candidate passages even though their keywords and structures are different (i.e., even having a lower surface similarity score).

5. Conclusion and future work

In this paper, we presented a new ontology for the Arabic NLP. The main objective of this ontology is filling in the gap registered in the availability of semantic Arabic resources. The proposed ontology combines the lexical information and hyponymy relations in AWN with the semantic and syntactic frames of verb classes in AVN.

In order to ensure the usability of this resource in a semantic-based application such as Arabic QA, we transformed the AVN frames into the CG formalism that allows the representation of meaning in questions and passages with the aim to compare both representations. Also, in this work, we presented experiments showing the promising coverage of our ontology with respect to state-of-art question test-sets (i.e., CLEF 2013). Consequently, the improvement of QA performance was registered either in terms of percentage of questions for which the system was able to provide answers as well as in terms of percentage of correctly answered questions that is significantly increased.

As future work, we plan to integrate the semantic reasoning approach based on the ontology we constructed as part of a multi-level Arabic QA system called “IDRAAQ” [5] as part of the Arabic NLP integrated platform “SAFAR” [30]. To improve the intelligent aspect of IDRAAQ, richer CGs will be considered with the coverage of different challenging questions and situations.

Acknowledgment

The work of the last author was carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP 7 Marie Curie, the DIANA APPLICATIONS - Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project, and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

Special thanks go to Jaouad Mousser, of University of Konstanz, Department of Linguistics, for allowing us to use the Arabic VerbNet resource.

References

[1] L. Abouenour, K. Bouzoubaa and P. Rosso, Structure-based evaluation of an Arabic semantic query expansion using the JIRS passage retrieval system. In: *Proceedings of the workshop on computational approaches to Semitic languages, E-ACL* (2009a), Athens, Greece, March.

[2] L. Abouenour, K. Bouzoubaa and P. Rosso, Three-level approach for passage retrieval in Arabic question /answering systems. In: *Proceedings of the 3rd international conference on Arabic language processing CITALA'09* (2009b), Rabat, Morocco, May.

[3] L. Abouenour, K. Bouzoubaa and P. Rosso, Using the YAGO ontology as a resource for the enrichment of named entities in Arabic WordNet. In *Workshop LR & HLT for semitic languages, LREC* (2010a). Malta. May.

[4] L. Abouenour, K. Bouzoubaa and P. Rosso, An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. *Special Issue in the International Journal on Information and Communication Technologies/IEEE* (2010b). June.

[5] L. Abouenour, K. Bouzoubaa and P. Rosso, IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval. In: *CLEF* (2012), Online Working Notes/Labs/Workshop.

[6] L. Abouenour, K. Bouzoubaa and P. Rosso, On the Evaluation and Improvement of Arabic WordNet Cov-

erage and Usability. In: *Languages Resources and Evaluation*, vol. 47, issue 3 (2013), pp. 891-917, June.

[7] M. Attia, M. Rashwan, M. Al-Badrashiny, Fassieh (R), A semi-automatic visual interactive tool for morphological, pos-tags, phonetic, and semantic annotation of arabic text corpora. In: *IEEE Trans Audio Speech Lang Process* 17 (2009), 916–925.

[8] O. Babko-Malaya, M. Palmer, N. Xue, A. Joshi, S. Kulick, Proposition Bank II: Delving deeper, frontiers in corpus annotation. In: *Workshop in conjunction with HLT/NAACL* (2004), Boston, MA, May 6.

[9] M. Diab, Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. In: *Proceedings of the 2nd Intl. Conference on Arabic Language Resources and Tools* (2009), Cairo, Egypt.

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

[11] O. Etzioni, M. J. Cafarella, D. Downey, S. Kok, A. M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, Web-scale information extraction in KnowItAll. In: *WWW* (2004).

[12] C. Fellbaum (ed.), WordNet – An Electronic Lexical Database. The MIT Press, Cambridge, 1998.

[13] S. Grimm, A. Abecker, J. Völker and S. Rudi, Ontologies and the Semantic Web. In *J. Domingue, D. Fensel, & J. A. Hendler, Handbook of Semantic Web Technologies,* 2011, S. 508-537, Berlin, Heidelberg: Springer.

[14] T. Gruber, A translation approach to portable ontology specifications. In: *Knowledge Acquisition* (1993), 5, 2, 199–220.

[15] N. Habash, O. Rambow, R. Roth, MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In: *Proceedings of the 2nd Intl. Conference on Arabic Language Resources and Tools* (2009), Cairo, Egypt.

[16] S. Hensman and J. Dunnion, Automatically building conceptual graphs using VerbNet and WordNet. In: *Proceeding of the 3rd International Symposium on Information and Communication Technologies* (2004), pp 115–120, Las Vegas, NV.

[17] Kabbaj, Development of Intelligent Systems and Multi-Agents Systems with Amine Platform. In: *Proceeding of the 15th Int. Conference on Conceptual Structures, ICCS* (2006), Springer-Verlag.

[18] Kabbaj, An Overview of Amine. In: *P. Hitzler and H. Scharfe (eds.), Conceptual Structures in Practice*, CRC Press, Taylor & Francis Group, 2009, pp. 321-347.

[19] K. Kipper-Schuler, VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, PA, 2005.

[20] Levin, *English Verb Classes And Alternations: A Preliminary Investigation*. The University of Chicago Press, 1993.

[21] Matuszek, J. Cabral, M. Witbrock, J. Deoliveira, An introduction to the syntax and content of cyc. In: *Proceedings of the AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and its Applications to Knowledge Representation and Question Answering* (2006).

[22] M. Montes-y-Gómez, A. Gelbukh, A. López-López, R. Baeza-Yates, Flexible Comparison of Conceptual Graphs. In: *Proceeding of the 12th International Conference on Database and Expert Systems Applications DEXA* (2001). *Lecture Notes in Computer Science*, vol 2113, Springer-Verlag, Munich, Germany, September.

[23] J. Mousser, A Large Coverage Verb Lexicon For Arabic. In: *Proceedings of the 7th conference on International Language Resources and Evaluation* (LREC) (2010), Valetta, Malta.

[24] J. Mousser, Classifying Arabic Verbs Using Sibling Classes. In: *Proceeding of the International Conference on Computational Semantics (IWCS)* (2011), Oxford, UK.

[25] M. T. Pazienza, M. Pennacchiotti and F. M. Zanzotto, Mixing WordNet, VerbNet and Propbank for studying verb relations. In: *Proceedings of the 5th Language Resource and Evaluation Conference LREC* (2006), Genoa, Italy.

[26] M. Kaisser, Web Question Answering by Exploiting Wide-Coverage Lexical Resources. In: *Proceedings of the 11th ESSLLI Student Session* (2006). J. Huitink & S. Katrenko (eds.).

[27] M. Moens and M. Steedman, Temporal ontology and temporal reference. In: *Computational Linguistics* (1988) 14, 15–28.

[28] Niles and A. Pease, Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: *Proceedings of the international conference on information and knowledge engineering* (2003), Las Vegas.

[29] M. Rashwan, M. Al-Badrashiny, M. Attia, S. Abdou, and A. Rafea, A stochastic Arabic diacritizer based on a hybrid of factorized and unfactorized textual features. In: *IEEE Transactions on Audio, Speech and Language Processing* (2011), vol. 19, no. 1, pp. 166-175.

[30] S. Sidrine, Y. Souteh, K. Bouzoubaa and T. Loukili, SAFAR: vers une Plateforme Ouverte pour le Traitement Automatique de la Langue Arabe. In: *Proceeding of the 6th Conference of Intelligent Systems: Theory and Applications SITA* (2010), May, Rabat, Morocco.

[31] F. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*, 1984, Addison-Wesley Company.

[32] P. Vossen (ed)., EuroWordNet, *A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1999, The Netherlands.

[33] R. Sutcliffe, A. Peñas, E. Hovy, P. Forner, A. Rodrigo and C. Forascu, Overview of QA4MRE Main Task, *CLEF* (2013).

[34] M. Shaheen and A. M. Ezzeldin, Arabic Question Answering: Systems, Resources, Tools, and Future Trends. In: *Arabian Journal for Science and Engineering* (2014), pp 1-24, Springer Berlin Heidelberg.

[35] Y. Benajiba, P. Rosso, L. Abouenour, O. Trigui, K. Bouzoubaa and L. H. Belguith, Question Answering, Chapter 11 in: *Natural Language Processing of Semitic Languages.* Zitouni I. (Ed.), 2014, Series: Theory and Applications of Natural Language Processing. Springer.