Ontology-Based Semantic Annotation of Arabic Language Text

Maha Al-Yahya, Mona Al-Shaman, Nehal Al-Otaiby, Wafa Al-Sultan, Asma Al-Zahrani, Mesheal Al-Dalbahie

Information Technology Department, College of Computer & Information Sciences, King Saud University
Riyadh, Saudi Arabia
Email: malyahya, 429200187, 429200199, 429202934, 429200194, 429205127 @ksu.edu.sa

Abstract—Semantic annotation is the process of adding semantic metadata to resources. Semantic metadata is data concerning the meaning of entities and the relationships that exist. Semantic annotation cannot be performed without an ontology suitable for the task. In this research paper, we describe the design, implementation, and evaluation of a lexical ontology for Arabic semantic relations. The main purpose of the ontology is to facilitate the task of semantic annotation of the Arabic textual content. The ontology was evaluated for usability and usefulness using a prototype system for the automated semantic annotation of Arabic text. The results of the evaluation indicated that the ontology was fit for the purpose of semantic annotation of Arabic text with lexical relations. The evaluation has also revealed important findings and recommendations for designing Arabic semantic annotation tools.

Index Terms—Ontology, semantic annotation, Arabic language, lexical semantic relation.

I. INTRODUCTION

The semantic web vision can only be realized when web resources, which are mainly composed of textual documents, are annotated with metadata that provides data meaning and enables machines to process information in intelligent forms. The process of associating metadata with resources is called annotation, and semantic annotation is the process of annotating resources with semantic metadata [1]. Semantic annotation serves multiple purposes and can be useful in many semantic web applications and natural language processing (NLP) tools. For example, web documents and textual corpora can be accurately searched, producing results that can be automatically aggregated by web services.

Generating semantic annotations manually is not feasible due to the sizeable amount of Arabic web content that is available today and that is continually increasing as a result of the increase in Internet users from the Arab world. According to the world Internet usage statistics [2], the Middle East has seen a tremendous increase in the number of users from 2000 to 2014. The growth rate is at 3,303.8%, which is the second highest among world continents.

The majority of semantic annotation tools available only support Latin languages, and adaptation to Arabic is complex [3]. Moreover, the amount of research in the area of Arabic content annotation is limited, and one of the main reasons is the lack of resources, such as ontologies. Semantic annotation cannot be performed without an ontology to map instances with concepts and relations. In this paper, we describe the design, implementation and evaluation of the SemTree ontology for lexical semantic relations for the task of semantic annotation of Arabic textual content. The ontology is tested using a prototype system that serves as a platform for automatically extracting relations and providing annotations.

The remainder of the paper is organized as follows: Section 2 presents related work on Arabic ontologies and Arabic semantic annotation tools. Section 3 provides an overview of the ontology design and development process. Section 4 describes the evaluation of the usefulness and usability of the SemTree ontology for the annotation task using the prototype system. Section 5 presents our conclusions and further research.

II. RELATED WORK

In the context of the semantic web, an ontology is a vital element in any semantic annotation task. The work presented in [4] describes how an ontology can be useful for information retrieval. Another example is the work presented in [5], which describes a model for semantic indexing of web documents using ontological concepts.

Work on Arabic ontologies can be described based on the purpose or objective for building the ontology, and the competency questions. A number of Arabic ontologies have been developed for improving the task of information retrieval. The work presented in [6] provides an Arabic ontology of legal terms to improve both recall and precision of Arabic information retrieval in the legal domain. The ontology is used in the process of query-expansion to improve precision. The ontology cannot be used across domains, as it is domain specific, and is not suitable for lexical semantic annotations.

Several Arabic ontologies have been built for the task of Arabic NLP. For example, the Al-Khalil ontology [7] is an Arabic ontology developed for Arabic linguists and provides detailed formal specifications of Arabic

language grammar to aid in Arabic NLP tasks. The approach they adopted for building the ontology is based on re-using existing linguistic ontologies, such as GOLD [8]. Similarly, in [9] and [10], the authors presented a lexicon, which formalizes the morphological knowledge of Arabic language verbs and their derivatives using an ontology. The main purpose of such formalism is to support Arabic NLP applications. They presented a model for the Arabic language by identifying the mathematical model and the derivation rules used for the ontology construction. The ontology is designed to model derivational rules for verbs in the Arabic language and does not include formalisms for modeling the relationships between words. In [11], Jarrar presented a proposal and methodology for building a formal Arabic ontology. The proposal was concerned with semantic relations between concepts instead of words. Relations include subtype, inheritance, hyponymy, and general-specific relations. The top level of the ontology is based on DOLCE [12] and SUMO [13] upper level ontologies. This ontology is still a work-in-progress, and our work could be an excellent resource for the Arabic Ontology project by utilizing the set of individuals in the SemTree to aid in the development of the ontology concepts and word glosses.

The Arabic Wikipedia has also been used to build ontologies and extract relations. For example, Al-Rajebah et al. [14] proposed a methodology for identifying ontology instances in which the Arabic version of Wikipedia is used as a knowledge source from which concepts and semantic relations are extracted. Their algorithm extracts the semantic relations between the article and the features it contains using Wikipedia "Infoboxes."

The Quranic Arabic Corpus ontology [15] provides an ontological classification of the concepts found in the Holy Quran. Searching for the word "dawn" in the ontology returns the subtree (concept/event/physical event/dawn), the verse in which it is cited, along with its translation, and a visual concept map. Although it provides higher-level conceptualizations, it does not address the meaning of specific words in the Holy Quran.

The Azhary [16] ontology is a lexical ontology for the Arabic language. Its primary use is in automatic text analysis and artificial intelligence applications. The Azhary system architecture contains the following three main elements: word extraction, relation building, and ontology building. They define seven types of lexical relations, and words are extracted from several lexicons.

The work presented in [17] describes an approach of automatic construction on an Arabic linguistic ontology using statistical techniques to extract elements of ontology from Arabic texts. They use the "repeated segment" technique to identify the relevant terms that represent main concepts in the domain. They also use "co-occurrence" of these concepts to define relations in the ontology.

With regard to the task of Arabic semantic annotation, one of the earliest tools is the AraTation tool. It was designed to annotate Arabic news documents on the Web.

The tool is capable of extracting named entities using an Arabic location ontology designed specifically for this purpose. Although AraTation uses an ontology for the annotation process, the ontology only covers location names and is not useful for lexical semantic relations. Similarly, the work presented in [18] describes a semantic annotation tool for the Arabic language, which is tested on a news article corpus. The system uses hard-coded rules to identify "Reported Information (RI)" and annotate them accordingly. The task includes a semantic annotation by specifying the author of the information, his/her position in relation to his/her remarks, and where or when he/she said them. It is focused on the task of automated RI location in text and annotation, and is not suited for lexical semantic relation annotation. Another tool that supports the semantic annotation of Arabic web documents is presented in [3]. The tool reads a URL and an ontology, and then it automatically generates annotations based on the words in the web page. It allocates instances in the text and annotates them with ontology vocabulary. The main objective is to enable semantic search engines to achieve higher recall and precision. The annotation tool was evaluated on the following three domain ontologies: food, nutrition, and health. The annotation tool is composed of two major components: the analyzer, which parses the web documents and extracts entities, and the matcher, which is responsible for locating matches between document entities and ontology terms. Once a match is found, the document is annotated using the annotator component. An important aspect of the tool is that it enables the provision of an arbitrary ontology; however, one of the limitations is that it is domain specific. Although the method is promising, it suffers from the limitation that only instance data, which are available in the ontology, are annotated in the document.

The survey on Arabic language support for semantic web tools and applications presented in [19] highlighted key areas for Arabic language support in semantic web tools and proposed a working framework. The framework is composed of two major components: ontology learning and a semantic query engine. Ontology learning involves extraction of Arabic terms, understanding meaning and linguistic relations between words, taxonomy classification, discovering other relations, and ontology alignment. Our work fits within the first two phases of the framework. In addition, a recent survey [20] by Al-Zoghbya, et al. provides an in-depth review on current trends in Arabic language support for semantic web applications. The survey categorizes work into the following four main categories: (1) Arabic ontologies, (2) Arabic named entity extraction, (3) semantic representation of religious texts, and (4) Arabic semantic search engines. With regard to Arabic ontologies, the study highlights that there is an immanent need to develop Arabic ontologies to serve semantic web applications. Moreover, the study concludes that Arabic ontologies are rare, and those that exist are limited to a specific closed domain and thus cannot be used on a larger scale. These surveys justify the requirement to

develop Arabic ontologies and contribute to the development of semantic web tools and applications.

Our work fills a gap in literature and focuses on the task of facilitating annotation of Arabic text with lexical semantic relations. Using ontologies for semantic annotation of Arabic web resources can result in vastly improved and automated search capabilities, resource discoveries, and information analytics.

III. ONTOLOGY DESIGN & DEVELOPMENT

The design and development of the ontology was accomplished using the ontology development methodology proposed by Uschold and King [21], shown in Fig. 1. To build an ontology, the following processes must be performed:

*   Identify the purpose and scope of the ontology: during this phase of ontology development, justification for building the ontology is presented and the intended uses are identified.
*   Build the ontology: two activities are performed during this phase. First, ontology capture includes identifying key concepts and relationships in the domain of interest. Second, ontology coding involves two main tasks, identifying the actual terms used in the ontology and writing the code (using ontology editors).
*   Evaluate the ontology: evaluation involves a technical judgment of the ontology according to the requirements and specifications.
*   Document the ontology: setting naming conventions and guidelines for sharing, maintaining, and developing the ontology.
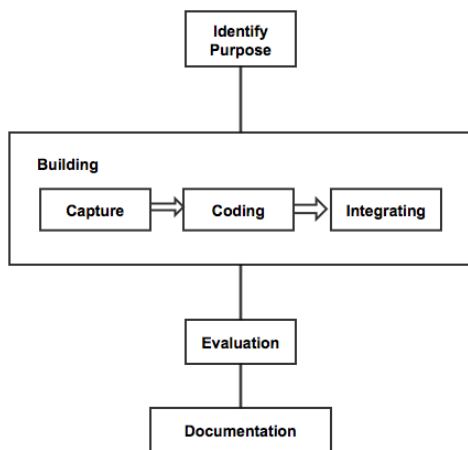

Fig.1. Ontology Design Methodology by Ushold & King [21].

A. Identify Purpose and Scope

An ontology can be built for various purposes (to be reused, shared, used as part of a knowledge base, as a vocabulary, etc.). For this research, the main purpose of the SemTree ontology is to provide a vocabulary for annotating Arabic language text with lexical semantic

relations. Such an ontology could be used for various purposes, including Arabic NLP tasks, information retrieval, or as a formal language lexicon. Concerning the scope, the ontology is to provide vocabulary and structure for the basic linguistic components for the Arabic language and lexical semantic relations. The major entities within the ontology are the following:

*   Linguistic units
*   Semantic fields
*   Semantic relations.

B. Ontology Capture

Ontology capture refers to the process of identifying key concepts and relationships in the domain and to produce precise and unambiguous textual definitions for these concepts and relationships [22]. In this section, we provide definitions for classes and individuals in the ontology.

Classes

The ontology consists of six classes of which there are three upper classes that are extracted from the SemQ ontology [23], [24]. Fig. 2 depicts the major classes in SemTree.

*   Linguistic-concept: a class that represents all the terminology used in the SemTree ontology.
*   Semantic-field: a class representing all existing semantic fields within a language (temporal, spatial, human, etc.).
*   Semantic-unit: a class representing a single word.
*   Word: a single distinct element of speech.
*   Letter (particle): a function word or a "letter", it does not provide meaning by itself; however, when it is combined with other words, it provides meaning.
*   Verb: a word that can take the form of a "verb", which describes an action or state of being.
*   Noun: a word that can take the form of a "noun", which is a word used to represent the names of people, places, or things.
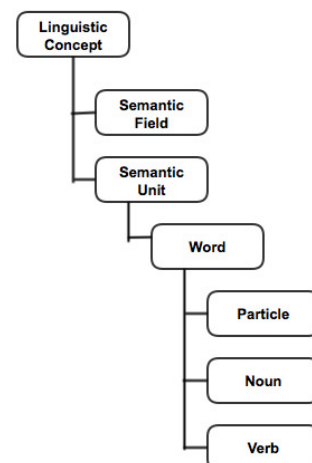

Fig.2. Major Classes in SemTree.

Individuals and Properties

Individuals in the ontology are of two types: words and semantic-fields. Properties are relations that exist between individuals in the ontology. Details of the properties are shown in Table I.

Table 1. SemTree Properties.

Property	Domain	Range	Type
isAntonymof	Words	Words	Symmetric
isSynonymof	Words	Words	Symmetric Transitive
hasSemanticField	Words	Words	Functional

Figs. 3 and 4 show a sample set of individuals from three semantic fields defined in the ontology. The diagram is read in the direction of the arrows. To simplify the diagram, not all properties are shown.
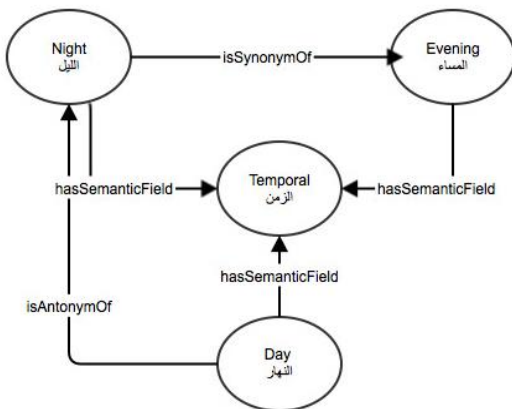
Fig.3. Sample Individuals from the Temporal Semantic Field.

C. Ontology Coding and Integration

Once the design was completed and all classes and properties were defined, we started coding the ontology using the Protégé ontology editor. The ontology was developed using the OWL/RDF language. Classes were defined and class hierarchies were created. Next, we created all properties in the ontologies with axioms.
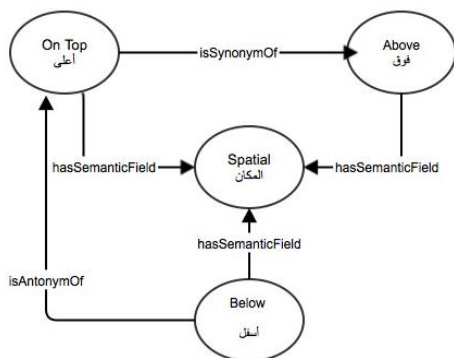

Fig.4. Sample Individuals from the Spatial Semantic Field.

All classes are considered subclasses of the class Thing, which is the default within the Protégé tool. Fig. 5 shows the class hierarchy in SemTree.
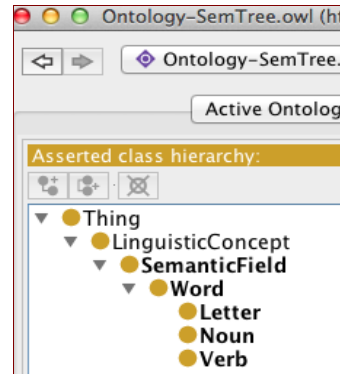

Fig.5. SemTree Class Hierarchy.

Next, we created the object properties, the relationships that exist between two individuals, as shown in Fig. 6. For each object property, we defined the range and domain of the property and all relevant property axioms.
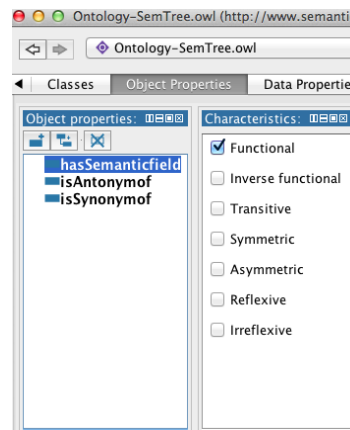

Fig.6. SemTree Properties.

Finally, the ontology was populated with sample individuals to test the classes and the properties. A sample of the individuals is shown in Fig. 7.
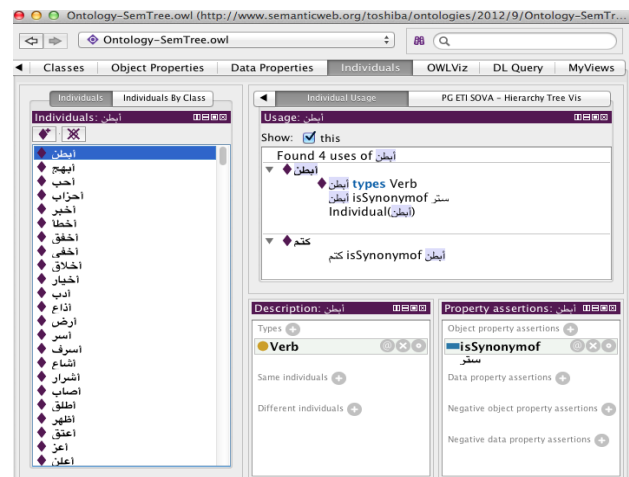

Fig.7. Sample SemTree Individuals.

IV. EVALUATION

Since the SemTree ontology is a task-oriented ontology and focuses on the practical use of the ontology for semantic annotations, our evaluation focuses on the usability and usefulness of the ontology in accomplishing the required task [25]. To evaluate the ontology, we designed a prototype system for automated lexical semantic annotations. The system was developed to support the process of ontology-based semantic annotation. It provided two major elements for semantic annotation: *entity identification* and *lexical relation extraction*.

For the entity identification element, the prototype system uses traditional shallow parsing to identify words in the text. This process includes word stemming and stop-word (low content, non-functional) removal.

For the relation extraction element, the system uses a set of manually predefined patterns for Arabic antonym and synonym relations and uses these patterns to identify new relations between words in the Arabic language text. Using regular expressions, the text is scanned for any matches. If a match is found, the lexical relation is extracted and the two words are annotated with the relevant relationship (isAntonymOf/isSynonymOf).

The user interface of the prototype system was designed to enable users to enter Arabic text freely or by uploading a file. Once the text is uploaded in the viewer window, the system enables the user to select which lexical relation to extract. The system then scans the text and annotates (highlights) all instances (words) that are a possible match for the chosen lexical relation.

The prototype system also enables the user to evaluate the accuracy of the lexical relations identified. If the relation extracted is correct, the new words are added to the SemTree ontology. The system creates individuals from the pair, creates the relationship, and adds the following to the SemTree ontology:

<http://www.ksu.edu.sa/Ontology/SemTree#Cold>
<http://www.ksu.edu.sa/Ontology/SemTree#isAntonymOf>
<http://www.ksu.edu.sa/Ontology/SemTree#Hot>.

This process is done for all correct lexical relations extracted.

The main user interface is in Arabic and is shown in Fig. 8. It is composed of three main windows, the *text entry/upload* window, *semantic relation selection* window, and *semantic relation display* window. The text entry/upload window enables the user to write text freely in the area or upload an available document. The semantic relation selection window enables the user to select the relation to extract from the given text. In the current implementation, the user may select an antonym relation, synonym relation, or both. The semantic relation display window allows the user to set preferences for viewing the results. These preferences are 'as highlighted in text' or 'as a list of words with corresponding relations'.
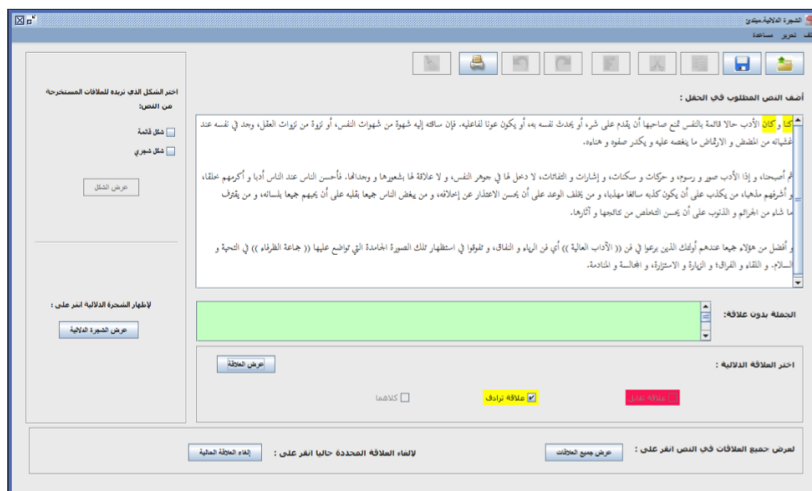

Fig.8. Lexical Semantic Relations Annotations Interface.

The user can also evaluate the semantic relations extracted by selecting the *evaluate relations* button. Fig. 9 shows the evaluation window. Here, all relations extracted from the text are displayed with two columns (correct and incorrect), and the user evaluates the accuracy of each item in the list.

The statistics collected from the evaluation assist in computing the accuracy of the lexical semantic relation extraction, and also ensure that the ontology establishes correct relations only.

Fig.9. Lexical Semantic Relation Evaluation.

The prototype system was tested by a number of users, and they were asked to provide feedback on their experience with the system. The users of the system did not face any problems while using the system, and the interface was self-explanatory and easy to use. However, the users provided several recommendations, which include (1) being able to see the ontology view in the editor with all classes as an expandable tree of relations and words, (2) enabling input of URLs, not just text files or free text in the text entry/upload window, and (3) viewing annotations in RDF/XML. Moreover, they highlighted the fact that the system was not able to extract and annotate all antonym and synonym relations in the text, and other semantic relations were extracted, such as hyponyms.

In addition, some of the relations annotated were incorrect; such information should be captured, and the system should be able to compute the accuracy of the annotation process.

From the evaluation, it is clear that accuracy statistics should be collected and measured. This will help in improving the system and increasing the accuracy of semantic annotations.

V. Conclusion and Further Reserach

In this paper, we presented the design, development, and evaluation of the SemTree ontology, an ontology for lexical semantic relation annotations for Arabic text. The usefulness of the ontology was evaluated using a prototype system that serves as a platform for extracting relations and providing annotations. The results from the evaluation indicated that the ontology was useful for the purpose of lexical semantic annotations. The evaluation also provided several recommendations for a lexical semantic annotation task. Further work on SemTree involves extending the ontology with other lexical semantic relations and enhancing the interface with annotation information, relation extraction accuracy computation, and ontology views. Moreover, the ontology and the tool can be used to annotate Arabic texts, thus developing a corpus for Arabic text annotated with semantic relations and used for corpus-based Arabic linguistics studies.

References

[1] M. Nagarajan, "Semantic annotations in web services," in *Semantic Web Services, Processes and Applications*, Springer, 2006, pp. 35–61.

[2] "World Internet Users Statistics and 2014 World Population Stats." [Online]. Available: http://www.internetworldstats.com/stats.htm. [Accessed: 05-Jan-2015].

[3] S. Al-Bukhitan, T. Helmy, and M. Al-Mulhem, "Semantic Annotation Tool for Annotating Arabic Web Documents," *Procedia Comput. Sci.*, vol. 32, pp. 429–436, 2014.

[4] V. Jain and M. Singh, "Ontology Based Information Retrieval in Semantic Web: A Survey," *Int. J. Inf. Technol. Comput. Sci.*, vol. 5, no. 10, pp. 62–69, Sep. 2013.

[5] A. Dennai and S. M. Benslimane, "Semantic Indexing of Web Documents Based on Domain Ontology," *Int. J. Inf. Technol. Comput. Sci.*, vol. 7, no. 2, pp. 1–11, Jan. 2015.

[6] S. Zaidi, M. Laskri, and K. Bechkoum, "A cross-language information retrieval based on an Arabic ontology in the legal domain," in *Proceedings of the International Conference on Signal-Image Technology and Internet-Based Systems (SITIS'05)*, 2005, pp. 86–91.

[7] Hassina Aliane, Zaia Alimazighi, and Ahmed Cherif Mazari, "Al - Khalil : The Arabic Linguistic Ontology Project," in *Proceedings of the International Conference on Language Resources and Evaluation*, Valletta, Malta, 2010.

[8] S. Farrar and D. T. Langendoen, "A linguistic ontology for the semantic web," *Glot Int.*, vol. 7, no. 3, pp. 97–100, 2003.

[9] Belkridem and El Sebai, "An Ontology Based Formalism for the Arabic Language Using Verbs and Derivatives," *Commun. IBIMA*, vol. 11, no. 5, pp. 44–52, 2009.

[10] F. Z. Belkredim and F. Meziane, "DEAR-ONTO: a derivational Arabic ontology based on verbs," *Int. J. Comput. Process. Lang.*, vol. 21, no. 03, pp. 279–291, 2008.

[11] M. Jarrar, "Building A Formal Arabic Ontology," in *Invited Paper). In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League. Tunis: April*, 2011, pp. 26–28.

[12] A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari, "Sweetening WORDNET with DOLCE," *AI Mag*, vol. 24, no. 3, pp. 13–24, 2003.

[13] "SUMO Knowledge base Browser." [Online]. Available: http://sigma.ontologyportal.org:4010/sigma/Browse.jsp?k b=SUMO&lang=en. [Accessed: 26-Jan-2011].

[14] N. I. Al-Rajebah, H. S. Al-Khalifa, and A. M. S. Al-Salman, "Exploiting Arabic Wikipedia for automatic ontology generation: A proposed approach," in *2011 International Conference on Semantic Technology and Information Retrieval (STAIR)*, 2011, pp. 70–76.

[15] K. Dukes, E. Atwell, and N. Habash, "Supervised collaboration for syntactic annotation of Quranic Arabic," *Lang. Resour. Eval.*, vol. 47, no. 1, pp. 33–62, Nov. 2011.

[16] H. Ishkewy, H. Harb, and H. Farhat, "Azhary: An Arabic Lexical Ontology," *Int. J. Web Semantic Technol.*, vol. 5, no. 4, pp. 71–82, Oct. 2014.

[17] Ahmed Cherif Mazari, Hassina Aliane, and Zaia Alimazighi, "Automatic construction of ontology from Arabic texts," presented at the International Conference Web and Information Technology 2012, Algeria, 2012.

[18] M. Alrahabi, A. H. Ibrahim, and J. -P. Desclés, "Semantic Annotation of Reported Information in Arabic," in *Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, Melbourne Beach, Florida, USA, May 11-13, 2006*, 2006, pp. 263–268.

[19] M. Beseiso, A. R. Ahmad, and R. Ismail, "An Arabic language framework for semantic web," in *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on*, 2011, pp. 7–11.

[20] A. M. Al-Zoghby, A. S. E. Ahmed, and T. T. Hamza, "Arabic Semantic Web Applications – A Survey," *J. Emerg. Technol. Web Intell.*, vol. 5, no. 1, Feb. 2013.

[21] M. Uschold and M. King, *Towards a methodology for building ontologies*. Citeseer, 1995.

[22] A. Gomez-Perez, O. Corcho, and M. Fernandez-Lopez, *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition*. London ; New York: Springer, 2004.

[23] M. Al-Yahya, H. Al_Khalifa, A. Bahanshal, I. Al-Odah, and N. Al-Helwah, "An Ontological Model for Representing Semantic Lexicons: An Application on Time Nouns in the Holy Quran," *Arab. J. Sci. Eng. AJSE*, vol. 35, no. 2C, pp. 21–35, 2010.

[24] H. S. Al-Khalifa, M. M. Al-Yahya, A. Bahanshal, and I. Al-Odah, "SemQ: A proposed framework for representing semantic opposition in the Holy Quran using Semantic Web technologies," in *2009 International Conference on the Current Trends in Information Technology (CTIT)*, Dubai, United Arab Emirates, 2009, pp. 1–4.

[25] J. Pak and L. Zhou, "A framework for ontology evaluation," in *Exploring the Grand Challenges for Next Generation E-Business*, Springer, 2011, pp. 10–18.

Authors' Profiles

Maha M. Al-Yahya is an Assistant Professor at the Information Technology Department, College of Computer and Information Sciences, King Saud University (KSU). She obtained her Masters degree in Computer Science from the University of Bristol, UK, and a PhD in Computer Science from the University of Nottingham, UK. She is currently the Head of Information Technology Department at the College of Computer and Information Sciences at King Saud University (KSU), Riyadh, Saudi Arabia. She is a member of the IWAN Research Group, which specializes in Web Technologies at the College of Computer and Information Sciences at KSU. Her research interests include the semantic web, ontological engineering, and learning technologies. She has published over 20 papers in international conferences and refereed journals.

Mona Al-Shaman, Nehal Al-Otaiby, Wafa Al-Sultan, Asma Al-Zahrani, and Mesheal Al-Dalbahie are students from the Information technology department, College of Computer and Information Sciences at King Saud University.