Journal of Informatics and Mathematical Sciences
Vol. 9, No. 4, pp. 1309–1314, 2017
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications

http://www.rgnpublications.com

Review Article

Ontology-based Semantic Representation for Arabic Text – a Survey

Mamdouh Farouk

Department of Computer Science, Assiut University, Egypt
*__Corresponding author:__ mamdouh@fci.au.edu.eg

Abstract. The explosion of information available on the Internet motivates researchers to semantically manipulate this information to enable Internet users to find what they actually need easily. One of the important pillars to manipulate data semantically is ontology. Arabic ontology is recently gains a lot of attention. This paper survey some available Arabic ontologies and compares between these ontologies to clarify the difference between the main categories of Arabic ontologies. Ontology creation approach and representation method is considered in this study.

Keywords. Semantic web; Arabic ontology; semantic representation; NLP

MSC. 18C50

Received: July 8, 2016 **Accepted:** December 17, 2016

1. Introduction

Since emerging semantic web vision, a lot of works have been introduced to enables computer to understand web contents. A lot of attentions have paid to semantically represent the web data based on predefined ontologies so that computer can understand and interact with web content in a smart way. The ontology is conceptualization of a specific domain [1]. On the other hand, some ontologies are domain independent. These ontologies are called general ontologies and can be used in different domains. One of the famous ontologies of English terms is WordNet.

Furthermore, different representations have been proposed for representing ontologies. Many of these representations are based on XML such as RDF (Resource Description Framework) [2]. However, WordNet is represented into a database schema.

Hundreds of millions of people are speaking Arabic language. It is the mother tongue of more than 20 countries. The research work in this area is still not enough [3]. Moreover, a hug amount of Arabic heritage is recently digitalized. Moreover, there is an explosion of Arabic data in internet. In order to make the use of these data, semantic methodology should be available to process this huge amount of Arabic data. One of the main steps towards processing Arabic data semantically is to represent it based on ontology. Consequently, creating Arabic ontologies is an essential task. Although there are many ontologies of English language and other languages, a few of Arabic ontologies have been introduces recently. This paper surveys the available Arabic ontologies and their features.

There are two main taxonomies for Arabic ontologies. the classifies Arabic ontologies according to building methodology. Ontologies are classified to automatic created, semi-automatic, and manually created ontologies. The second classifies Arabic ontologies depending on the purposed or domain of the ontology. Ontologies are classified to islamic purpose ontology, domain specific ontology, and general ontologies.

The rest of the paper is organized as follow: section two gives a brief description of some Arabic ontologies. The comparison between different Arabic ontologies is stated in section three. Finally, the conclusion of this work is stated in section four.

2. Arabic Ontologies

There are many Arabic ontologies are available on the Internet. Each ontology has its own characteristics. We can classify Arabic ontologies into two main categories. The first is general ontology and the second is domain specific ontologies.

In the following subsections, some Arabic ontologies are explained. For each ontology, we describe the ontology structure, the building method, number of terms and relations used in this ontology.

3. General Arabic Ontologies

The main purpose of creating general ontology is to be used in representing general text such as web pages, news paper, or text book. Moreover, natural language processing techniques are involved with this type of ontologies in manipulating Arabic text semantically.

The following subsections describe some general Arabic ontologies in details.

A. Azhary

Azhary is an Arabic lexical ontology that constructed on the same way as WordNet ontology. It gathers Arabic words into groups called synsets. Each synset contains a set of words that have the same meaning. So, a synset represents a single concept. Moreover, it establishes relations between these synsets.

It contains more than 26000 word organized in 13328 synset [3]. The ontology building phases contains: word extraction form quran corpus. The second phase is building relations between the extracted words. The relation in this phase is manually created. There are 7 types of relations used in Azhary ontology (synonym, hypernym, hyponym, meronym, holonym,

antonym, and association). The next phase is represented these synsets and relations into RDF format.

B. Arabic WordNet

AWN is constructed according to the methods developed for EuroWordNet which applied to many languages [4]. Like other WordNet thesauruses AWN is represented into a database format. It classifies Arabic words into synsets which are manually validated. It provides a deep semantic underpinning for each concept. A mapping to an upper ontology (SUMO) and bilingual dictionaries are used to construct the terms of AWN [5]. Furthermore, a manual refinement process is done to enrich the resulted synsets.

C. UNL

UNL was initiated in the Institute of Advance Studies of United Nation University in Tokyo. Its main objective is to capture the meaning of written documents. Originally, UNL started to solve the problem of language diversity on the Internet. This means that the user can search the web in any language and the results may contain resources which have been written in different natural languages. UNL is based on the representation of concepts and the relations between concepts. Furthermore, UNL can be used to represent knowledge of different linguistic levels. The UNL representation of text can be used for intelligent natural language processing [**?**].

UNL is one of the three forms (RDF, UNL, CDL) that can be used to express Common Web Language (CWL), which is part of the incubator activity of W3C [3]. CWL is a common language for exchanging information through the web and also for enabling computers to process information semantically. It is designed to replicate the natural language.

UNL represents a sentence as a semantic network in which notes are concepts and arcs are semantic relations between concepts. Furthermore, UNL can be used to describe all information and knowledge conveyed by natural language [**?**]. UNL describes the concept structure of the text based on a set of predefined semantic relations.

The design principles are different in UNL and RDF. RDF is designed to describe properties of resources, whereas UNL is designed to describe meaning of written documents regardless the used language. The main advantage of using UNL is that it does not depend on domain ontologies. However, it depends on a single ontology called Universal Words (UWs), which provides the vocabulary for UNL and the semantic background of each concept. UNL ontology contains terms from more than 40 different languages. Moreover, it contains large number of entries in each language. For example, it contains more than two million entries for English language according to UNL website, www.undl.org.

In UNL dictionary, a word is represented as a Universal Word in this form headword(constraint list). The same word may have different constraint part according to the different meaning of this word. For example, according to UNL ontology, there are different meanings for the word state

Entries in UNL dictionary are organized in is-a like hierarchy in which each word is connected to a set of related words through UNL relations such as icl (inclusion), iof (instance of), equ (equivalent to). Some words may have more than one super-class depending on word

meaning. The word is defined by a set of relations that relate the word with other words. These set of relations determine the real meaning and the behavior of this concept. Therefore, the UNL dictionary provides enough information to understand semantic structure of a sentence.

4. Domain Specific Arabic Ontologies

Domain specific ontology is created to cover concepts which belong to specific field of knowledge. For example, Quran ontology is a domain ontology which tries to cover the concepts existing in Holy Quran textbook. Moreover, this type of ontology is used to manipulate data in this domain only and it fails in other domains. Unlike general ontology, usually this kind of ontology is represented in semantic web languages such as RDF.

A. Quran Ontology (leads university)

A lot of research have focused on semantic application of quranic data. Semantic Search of Qur'an text have gain special attention [6]. In order to accomplish the task of semantic search on Qur'an, ontology for quranic concepts have been proposed. One of the most famous and robust Qur'an ontology is Qur'an corpus which developed in leads University. It was built using protégé tool for building ontologies. It is focuses on quranic concepts.

B. Hadith ontology

It is domain ontology in Hadith field. It tries to represent prophet Mohamed Hadith and its related issues specially the narration chain [7]. It contains a limited number of concepts (6 concepts). However, it contains a large number of individuals. This ontology can be used to suggest a judgment for a hadith depending on its narration chain.

5. Comparison between Arabic Ontologies

This section compares between different Arabic ontologies which are available to use. Different dimensions are considered in this discussion such as: number of concepts included in the ontology, whether the ontology is general or domain specific, creating approach, whether it depends on words or concepts, number of used relations, types of relations, representation language.

There are two types of ontologies: the domain specific ontology and general ontology. The general ontology target is to contain a large number of Arabic words and the relation between these words. The relations between words is lexical relations (such as synonym, hyponym, ...) The general ontology is normally used for manipulating Arabic text semantically or for natural language processing. On the other hand, domain specific ontology is built for containing concepts not words and the relations between these concepts are user-defined semantic relations. This means that the relations are different and depend on concerns of the user (ontology creator) and what he wants to represent on the ontology. The target of building domain specific ontology is to represent a specific corpus or resources. Usually, the domain specific ontology is represented in a semantic web language such as RDF.

Table 1. Comparison between Arabic ontologies

Criteria	Azhary	Arabic WordNet	UNL	Quran ontology	Hadith ontology
Number of words	26195	23496			
Number of Concepts/synsets	13328	11270	~200000 entries	46	6
Number of instances				110802	> 30000
General/domain	general	general	general	domain	domain
Construction	semi-automatic	semi-automatic	manual		manual
Number of relation types	7	3	46		37
Representation	RDF	database	UNL		RDF (protégé)

6. Conclusion

Building Arabic ontology is very important to manipulate the continuously growing Arabic documents in the Internet. There are some available Arabic ontologies that can be classified to general ontology and domain specific. Many factors are including in ontology creation and representation. This paper tries to clarify the difference between these ontologies to help the reader to select the proper ontology according to his application.

Competing Interests

The author declares that he has no competing interests.

Authors' Contributions

The author wrote, read and approved the final manuscript.

References

[1] G. Guizzardi, On Ontology, ontologies, Conceptualizations, Modeling Languages and (Meta) Models, in O. Vasilecas, J. Edler and A. Caplinskas (eds.), Frontiers in Artificial Intelligence and Applications, Databases and Information Systems IV, IOS Press, Amsterdam (2007).

[2] M. Farouk and M. Ishizuka, CDL-Based Semantic Representation For Dynamic Web Pages, *International Journal of Semantic Computing* **6**(1), 51 – 65, 2012.

[3] H. Ishkewy, H. Harb and H. Farahat, Azhary: An Arabic Lexical Ontology, Azhar University, Faculty of Engineering, Computers and Systems Engineering Department, *International Journal of Web & Semantic Technology* **5**(4), 2014.

[4] C. Fellbaum, M. Alkhalifa, W. J. Black, S. Elkateb, A. Pease, H. Rodríguez and P. Vossen (2006), Introducing the Arabic WordNet project, Proceedings of the 3rd Global Wordnet Conference, Jeju Island, Korea, January, 2006.

[5] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. A. Martí, W. Black, S. Elkateb, J. Kirk, A. Pease, P. Vossen and C. Fellbaum, Arabic WordNet: Current State and Future Extensions, in Proceedings of the Fourth International GlobalWordNet Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008.

[6] H. U. Khan, S. M. Saqlain, M. Shoaib, M. Sher, Ontology Based Semantic Search in Holy Quran, *International Journal of Future Computer and Communication* **2**(6), 2013.

[7] R. S. Baraka and Y. M. Dalloul, Building Hadith Ontology to Support the Authenticity of Island, *International Journal on Islamic Applications in Computer Science and Technology* **2**(1), 2014.