Towards Ontology Construction from Arabic Texts-
A Proposed Framework

Abeer Al-Arfaj and AbdulMalik Al-Salman

Computer Sciences Department, College of Computer & Information Sciences
King Saud University
Riyadh, Saudi Arabia
abeerarfaj@yahoo.com, salman@ksu.edu.sa

Abstract—Ontology is one of the most popular representation model used for knowledge representation, sharing and reusing. In light of the importance of ontology, different methodologies for building ontologies have been proposed. Ontology construction is a difficult and time-consuming process. Many efforts have been made to help ontology engineers to construct ontologies and to overcome the bottleneck of knowledge acquisition. But the majority of these methods have focused on the English language. There is a huge amount of information and knowledge in Arabic documents. So, there is a dispirit need for building Arabic ontology. In this paper, we propose a framework for ontology construction from Arabic texts based on The Holy Hadith (Sayings of the Prophet Mohammed). We also summarize the state of the art in Arabic ontology construction. The challenges facing ontology engineers in constructing ontology from Arabic texts and the way these challenges might be solved is discussed.

Keywords—Ontology learning, Arabic Ontology, Arabic Language, Concept discovery, Semantic relation extraction.

I. INTRODUCTION

Ontology is a backbone technology for the Semantic Web. It plays an important role in supporting knowledge based applications in the Semantic Web. Since ontology is a representation model which defines domain knowledge with explicit specifications that solve interoperability between human and machine, therefore it is used for knowledge representation, sharing and reusing.

Ontology has been used in wide applications like knowledge management, information retrieval, information integration, bioinformatics and e-learning [44].

Due to the importance of ontologies in these areas, different methodologies for building ontologies have been proposed [29,38]. Nevertheless, the manual process is subjective, very hard, time-consuming, error-prone, and can cost very much. This process becomes the bottleneck of ontology acquisition. To overcome these difficulties, research area known as ontology learning is generated.

Ontology learning is defined as the set of methods and techniques used for building an ontology from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources. Ontology learning techniques rely on methods from various fields such as machine learning, knowledge acquisition, Natural Language Processing (NLP), statistics, and information retrieval. Such techniques facilitate and support the construction of ontologies by the ontology engineer. This is the reason why ontology learning frameworks have been developed in the last years and integrated with standard ontology engineering tools. Ontology learning can be applied to unstructured, semi-structured and fully structured data to support semi-automatic and cooperative ontology engineering [19,43,45].

The ontology term has been adopted from philosophy, where it is defined as the “theory of existence”. Ontology is a well-known term in the field of AI and knowledge engineering. The most popular definition of ontology in information technology and the AI community was made by Gruber [24], which states that: “An ontology is a formal, explicit specification of a shared conceptualization”. According to Studer et al. [40], conceptualization refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. Explicit means that the type of concepts used and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects that an ontology should capture consensual knowledge accepted by different communities.

Although there are many of researches done in this area, there is lacking of studies in Arabic ontology. Constructing Arabic ontology is attractive since Arabic is the language of the holy Quran and the official language of a significant amount of countries, and it has millions of speakers worldwide. Hence, there is a huge amount of information and knowledge in Arabic documents. So, extracting knowledge from such texts would be beneficial and of great help for the Arabic speaking community. The Arabic language compared with the English language has a much more complex syntax. So, the need for new methods to construct ontology from Arabic texts is growing. The Arabic ontology is necessary knowledge for applications that processing Arabic documents.
Contributions of this paper are as follows:

- We propose a framework for ontology construction from Arabic texts based on The Holy Hadith (Sayings of the Prophet Mohammed).
- We classify the existing Arabic ontology construction approaches.
- We discuss the challenges facing ontology engineers in constructing ontology from Arabic texts and the way these challenges might be solved.

The paper is organized as follows. After the introduction, characteristics of the Arabic language and the challenges that face ontology engineers in constructing ontology from Arabic text is discussed in Section 2. In Section 3, an existing works in Arabic ontology construction is reviewed. A framework for ontology construction from Arabic text is proposed in Section 5. Finally, Section 6 concludes the paper.

II. ARABIC ONTOLOGY CONSTRUCTION: CHALLENGES AND TRENDS

Arabic language is one of the most important languages because it is the language of the holy Quran and the official language of a significant amount of countries, and it has millions of speakers world-wide. It is a Semitic language of 29 sound alphabets (36 written alphabets), and one of the united nations official languages. The increasing amounts of information and knowledge in Arabic documents have generated a growing need to represent the content of these documents in structure model. Arabic ontology is the foundation of the creation of Semantic Web in Arabic language [11,12]. There are many challenges that face the ontology engineers when constructing ontology from Arabic texts. We need tools to analyze corpus in order to construct ontologies from Arabic texts. We have many Arabic morphological analyzers that have been successful in solving morphology related issues. For example, Xerox Arabic Morphological Analysis and Generation [9], Buckwalter Arabic Morphological Analyzer (BAMA) [17], Alkhailil Morpho Sys [18] and AraComLex [7]. Arabic syntax has also been addressed by some researchers and some success has been achieved. Only few works reported on the terminological and knowledge representation of Arabic. Unfortunately, there are no special corpora for Arabic domains to be used with the tools, which are used to build ontologies for domains. We need semantic resources (such as machine readable dictionaries) to understand the senses of words and expressions. Due to complexity of Arabic language, building domain terminological resources is difficult.

A. Characteristics of the Arabic Language

Arabic is written with optional diacritics that specify short vowels, they are usually absent in written Arabic, which contributes to ambiguity. It is composed of nouns, verbs and particles. Nouns and verbs are morphemes and derived from a closed set of around 10,000 roots. Particles are used to complete the meaning of verbs and nouns [13]. The Arabic language has complex morphological, grammatical, and semantic aspects since it is a highly inflectional and derivational language, which makes morphological analysis a very complex task. Therefore, the NLP tools that were designed for English cannot exactly meet the need of the Arabic language. In addition, the Arabic Language lacks the capitalization feature which makes the extraction of the Arabic Named Entities a complex task. The Arabic language is highly ambiguous when vowelization feature is absent [11-13,22]. Many levels of ambiguity pose a significant challenge to researchers developing NLP systems for Arabic [6]. Researchers have found ambiguity in Arabic to be present at several levels of analysis [16, 22]. Internal word structure ambiguity: That is, when a complex Arabic word could be segmented in different ways. Syntactic ambiguity, Semantic ambiguity, Constituent boundary ambiguity and Anaphoric ambiguity. All these difficulties hinder the construction of Arabic ontologies.

III. RELATED WORK ON ARABIC ONTOLOGY CONSTRUCTION

With regards to ontology construction for the Arabic language, there have been very few efforts and most of these adopted a manual or a semi-automatic approach. One of them is the work presented by Zaidi et.al.[42] to construct an Arabic ontology in the legal domain according to steps proposed by [34]. They have used a top-down strategy. To each concept they associate its synonyms together with a restriction of derivatives set, which is the most strongly related to the legal domain.

On the other hand, Elkateb et.al.[13] introduced Arabic WordNet which is lexical resource for Modern Standard Arabic based on the design and contents of the universally accepted Princeton WordNet (PWN) to enable translation on the lexical level to English and dozens of other languages. They extend the ontology and its set of mappings to provide formal terms and definitions equivalent to each synonym set. In another work by Belkridem & El Sebai [14] ontological representation for the Arabic Language is presented. The ontology was developed based on the rules that govern the language. They used the derivations and their patterns to structure the Arabic language and to strongly link the words’ morphology to their semantics. They evaluated their model by using it with some real application and comparing it with other traditional Approaches.

Also, a tool that generates an Arabic similarity thesaurus automatically from a text corpus is presented by Al-Qabbany et.al.[4]. They improved the similarity thesaurus construction method used for query expansion in information retrieval. Their results showed an improvement of about 3.3% over its predecessor method, to reach 98.1%.

Arabic ontology extraction from web documents is presented in Hazman et.al.[26]. They used both noun phrases appearing in the headings of a document and the document’s hierarchical structure to extract concepts and is-a relations between them in
a document’s domain. The ontology is constructed through the use of two complementary approaches. In the first approach (N-gram based), all phrases that start with the root concept in the document’s heading titles are extracted and considered instances of that concept. While the second approach uses the hierarchical structure as represented by HTML heading tags for discovering the children of the root concept in order to build its hierarchical ontology. They experimented their system by building an ontology in the agricultural domain using a set of Arabic documents. They used frequency measures to filter concepts and they evaluated the ontology against a modified version of AGROVOC ontology, which is a hand-made ontology. The F-score obtained was 52.29% for lexical evaluation of diseases ontology and 39.64% for lexical evaluation of insects’ ontology. Taxonomical F-score was 44.59% for diseases ontology and 31.38% for insects’ ontology.

In the Quranic Ontology [31], a knowledge representation is used to define the key concepts in the Quran, and to show the relationships among these concepts using predicate logic. Named entities in verses, such as the names of historic people and places mentioned in the Quran, are linked to concepts in the ontology as part of named entity tagging. The ontology also defines a set of semantic relations between these concepts. It uses NLP without taking into consideration the meaning of words in the Holy Quran.

Alyahya et.al.[1] proposed a computational model for representing Arabic lexicons using ontologies. The model is based on the field theory of semantics from the linguistics domain, and the data which drives the design of the model is obtained from the Holy Quran. They followed in the ontology development the UPON (Unified Process for ONtology) ontological engineering approach. The ontology was implemented using OWL (Web ontology language). From their evaluation results, they concluded that the model is capable of representing word semantics in a way that can facilitate semantic analysis of Arabic words and various useful applications.

Al-Khalil linguistic ontology by Aliane et.al.[5] is an OWL ontology which is based on the GOLD linguistic ontology. Their development approach of the ontology is in two steps: first, bootstrapping manually the ontology by choosing the linguistic concepts from Arabic linguistics and relating them to the concepts in GOLD. Second step is using an automatic extraction algorithm to extract new concepts from linguistic texts to enrich the ontology. No implementation is available for this project.

In another work, Alsafadi et.al.[8] introduced a domain-dependent Ontology for searching Arabic blogs in the Computer Technology domain. The authors proposed a model for designing the Ontology which is based on structuring the Arabic language into a set of equivalent classes, properties and relationships.

Recently Al-Rajebah & Al-Khalifa [3] proposed a model to extract ontologies from Wikipedia using a linguistic approach based on the semantic field theory introduced by Jost Trier. They applied the proposed approach on the Arabic version of Wikipedia. The semantic relations were extracted from infoboxes, hyperlinks within infoboxes and list of categories that articles belong to. To evaluate their system, they conducted three experiments which are: validity testing of the ontology according to OWL rules and human judgments from experts and the crowded. The system output achieved an average precision of 65%.

Jarrar [30] proposed to map knowledge extracted from machine readable dictionaries and linked to the concepts of SUMO and DOLCE with the English WordNet. This type of approach is limited, because we should guarantee the translation process and accuracy.

Annotating Web content in Arabic language has received less attention compared to Latin Languages. Albukhitan & Helmy[2] presented an automatic annotation of the Arabic Web resources related to food, nutrition and health domains. Their proposed method makes use of developed Arabic OWL ontologies related to those domains. It used linguistic patterns to discover relevant relationships between the named entities in the Arabic Web resources. The extracted information is then associated to the corresponding concepts and object properties of the developed ontology to produce the RDF metadata for the corresponding Web resources. Empirical evaluations of the proposed method show promising precision and recall.

Recently Harrag et al.[25] used association rules to extract the ontology of prophetic narrations (Hadith). They investigated the use of association rules to identify frequent item-sets over concepts that are related to Islamic jurisprudence (Fiqh) from the Sahih Al-Bukhari documents by computing correspondence relations using the Apriori algorithm. The association rules express any relation between two classes of connected concepts in the Sahih Al-Bukhari collection.

Beseiso et al.[11] conducted a survey of semantic Web technologies that support Arabic. They investigated four mostly used semantic Web tools namely Protégé, Jena, Sesame and KOAN with respect to their functionality, supported standards and degree for Arabic language support. They concluded that those systems do not completely support Arabic language processing or diacritics and thus, the need for new tools to be developed in supporting NLP for Arabic is crucial. Moreover, it is a must for development and design of semantic tools that support Arabic language processing & encoding.

Table1 shows a summary of some works on ontology construction from Arabic text. Based on the conducted studies, existing works on Arabic ontology can be classified into the following approaches: Manual approach, Statistical approach, Linguistic approach that uses linguistic knowledge to construct and enrich ontology, and Translation approach that uses knowledge already modeled in other languages to build a new resources for Arabic language. Translation approach is limited since it depends on languages linguistic environment and culture contexts.
TABLE 1. CLASSIFICATION OF ARABIC ONTOLOGY CONSTRUCTION APPROACHES

<table>
<thead>
<tr>
<th>Approach</th>
<th>Research</th>
<th>Learning resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>Zaidi et al., 2005 [42]</td>
<td>Unstructured text</td>
</tr>
<tr>
<td></td>
<td>Alsafadi et al., 2011 [8]</td>
<td></td>
</tr>
<tr>
<td>approach</td>
<td>Al-Rajebah &amp; Al-Khalifa 2013 [3]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Al Bukhitan &amp; Helmy 2013 [2]</td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td>Elkateb et al., 2006 [13]</td>
<td>Structured text</td>
</tr>
<tr>
<td>approach</td>
<td>Jarrar 2011 [30]</td>
<td></td>
</tr>
<tr>
<td>Statistical</td>
<td>Al Qabbany et al., 2009 [4]</td>
<td>Unstructured text</td>
</tr>
<tr>
<td>approach</td>
<td>Hazman et al., 2009 [26]</td>
<td>Semi-structured text</td>
</tr>
<tr>
<td></td>
<td>Harrag et al., 2013 [25]</td>
<td></td>
</tr>
</tbody>
</table>

IV. A FRAMEWORK FOR ONTOLOGY CONSTRUCTION FROM ARABIC TEXTS

Ontology construction is a complex task, including several steps as terminology extraction, recognition, finding relations between extracted terms and adding them in the building ontology, and in every step specific algorithms are needed. In this paper we propose a framework for ontology construction from The Holy Hadith (Sayings of the Prophet Mohammed) corpus considering the challenges and limited resources.

As shown in Figure 1, the input is the Holy Hadith (Al-Hadith) corpus, and the output is ontology constructed from Al-Hadith texts. There are four main phases: Preprocessing of corpus, Concept Extraction, Concept relation exploration and Ontology building.

A. Preprocessing of Corpus

The aim of this phase is to process Arabic text documents via content analysis by employing a variety of Arabic natural language processing techniques to make Arabic documents ready for extraction algorithms. In our work, we believe that the preprocessing of Arabic documents is a challenging and crucial stage. It may impact positively or negatively on the accuracy of ontology learning tasks. Therefore, the choice of preprocessing approaches will lead to improvement of ontology learning tasks.

B. Concept Extraction

To build Arabic ontology, the initial step is to find the important concepts of the target domain. As terms correspond to linguistic representation of concepts in the texts, concept extraction is thus to extract those domain specific terms from texts. For English there are some studies done for concept extraction, moreover, there are some studies for unstructured Arabic documents for key phrase extraction and multiword terms extraction such as [15,16,21,39,42]. However, key phrase extraction is different from concept extraction. In our framework, concept extraction consists of terminology extraction and concept identification. We need an algorithm to select relevant concepts for the ontology.

Figure 1. A Framework for Ontology Construction from Arabic Texts
The algorithm aims to group sets of candidates into a unique set of concepts. After that the result can be validated and further improved by expert to get a more refined list of concepts. This module plays a key role in the ontology learning process, whose performance greatly affects the system’s overall performance for building domain ontologies.

C. Relation Extraction

Semantic relations are an important component of ontologies that can support many applications e.g. text mining, question answering, and information extraction. However, the task of relation extraction is one of the main challenges in Natural Language Processing. We will propose a model for Relation Extraction from Arabic texts, which combines linguistic, statistic and data mining techniques to determine the relationships between concepts. There is no single step acquisition of knowledge about semantic relations. Rather, we claim, extracting semantic relations from Arabic text requires a model for integrating linguistic and non-linguistic knowledge. Relation extraction module includes three tasks: Extraction of candidate relations, Refinement and Evaluation.

D. Ontology Edition

In this phase, the ontology is constructed from the elements previously extracted with the purpose of detecting classes, subclasses and properties of the ontology. The ontology engineers will have the facility to edit the learned concepts and relations. Within this framework, the domain expert will make the final decision to select relevant concepts and relations. The selected results will be exported to the ontology in one of the ontology representation language (RDF, RDFS, OWL).

V. CONCLUSION

In this paper we have presented an overview of ontology learning. A definition of an ontology was provided, the tasks involved in the ontology learning process were identified and the most commonly cited approaches for this problem were introduced. We have briefly discussed the importance and characteristics of Arabic language. Further, We have also provided a brief overview of some works on ontology construction from Arabic text followed by a summarized comparison of them in Table1. We also discussed some of the current issues and open questions of the ontology construction from Arabic texts.

As ontologies are playing an important role in the Semantic Web, we can conclude that, we need to construct ontology from Arabic documents. Almost all systems built for ontology learning have been developed in languages other than Arabic. There are no systems or tools that support the construction of ontologies for Arabic language domains. In this paper we have proposed a framework to construct ontology from Arabic texts. This framework is based on NLP, statistical and data mining methods for extracting concepts and semantic relations. We believe that a significant challenge for Arabic ontology construction is the lack of systematic evaluation methods and reference standards. Thus, we need a corpora and ontological gold standards.

ACKNOWLEDGMENT

We would like to thank Dr. Fouzi Harrag for providing ideas and comments.

REFERENCES


