

Contents lists available at ScienceDirect



Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns

Rahima Bentrchia ^{a,*}, Samir Zidat ^a, Farhi Marir ^b^a Department of Computer Science, Lastic Laboratory of the Systems and the information and Communication Technologies, Chadid Mostefa Ben Boulaid, University of Batna 2, Batna, Algeria^b College of Technological Innovation, Zayed University, Dubai, United Arab Emirates

ARTICLE INFO

Article history:

Received 29 June 2017

Revised 13 August 2017

Accepted 11 September 2017

Available online xxx

Keywords:

Ontology

Text mining

Arabic AND conjunctive patterns

Arabic grammar

Semantic relations

Quranic Arabic Corpus

ABSTRACT

There is an immense need for information systems that rely on Arabic Quranic ontologies to provide a precise and comprehensive knowledge to the world. Since semantic relations are a vital component in any ontology and many applications in Natural Language Processing strongly depend on them, this motivates the development of our approach to extract semantic relations from the Quranic Arabic Corpus, written in Arabic script, and enrich the automatic construction of Quran ontology. We focus on semantic relations resulting from proposed conjunctive patterns which include two terms with the conjunctive AND enclosed in between. The strength of each relation is measured based on the correlation coefficient. Finally, we evaluate the significance of this method by using hypotheses testing and Student *t*-test. The obtained results are very promising since we combine an accurate Arabic grammar with strong statistical techniques to prove the existence and measure the strength of this type of semantic relations.

© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

An early definition of the term ontology appeared in 1993 (Gruber, 1993), where it is defined as a specification of a conceptualization. Maedche and Staab stated the original description of ontology learning from a text as the acquisition of a domain model from data (Maedche and Staab, 2001), where the extracted knowledge from the text is represented by concepts and relationships. Hence, semantic relations are an important element in the construction of ontologies (Alvarez et al., 2007). Besides holding together the concepts that represent the domain, they solve the ontology structuring problems. Furthermore, providing richer semantics to these relations facilitates selecting the operations that can be performed on them and the task that the ontology can tackle. However, semantic relations have not been given the attention they deserve because of the difficulty to capture the whole

information related to the problem domain as well as the different possible representations provided for a specific relation.

The process of ontology learning passes through several tasks organized in a layer cake. Each layer is explained deeply in (Cimiano, 2006; Liu et al., 2011). Traditionally, ontology construction depends on domain experts, but it is lengthy, costly, and controversial (Navigli et al., 2003). Therefore, automatic ontology construction approach was suggested but it is also still a difficult task due to the lack of a structured knowledge base or domain thesaurus (Lee et al., 2007).

On the other side, there is a lack in the developed approaches that deal with ontology learning from texts written in Arabic script due, for example, to the nature of Arabic writing, the semantic ambiguity of words, and the shortage in resources and tools that support Arabic (Farghaly and Shaalan, 2009). For Quran ontologies, all studies aim to achieve the purpose of understanding Quran as a source of knowledge and facilitating information retrieval automatically. Therefore, Quran can be presented to the world and employed very efficiently in many linguistic and religious studies. Currently, there are no complete Quran ontologies; many of them have covered specific topics in Quran or special types of words rather than the whole Quran words (Saad et al., 2010). Also, many researchers have built ontologies for parts of Quran and very few have used the entire Quran. Moreover, each ontology has focussed only on one or two types of relations between terms such as

* Corresponding author.

E-mail addresses: rahmabentrchia@yahoo.com (R. Bentrchia), samir.zidat@gmail.com (S. Zidat), Farhi.marir@zu.ac.ae (F. Marir).

Peer review under responsibility of King Saud University.



Production and hosting by Elsevier

<http://dx.doi.org/10.1016/j.jksuci.2017.09.004>

1319–1578/© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Please cite this article in press as: Bentrchia, R., et al. Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns. Journal of King Saud University – Computer and Information Sciences (2017), <http://dx.doi.org/10.1016/j.jksuci.2017.09.004>

synonymy and meronymy (Shoabib et al., 2009). As a validation method, the existing ontologies are evaluated by domain experts who rely on scholarly sources in their decisions (Ta'a et al., 2013).

Despite all these challenges, ontologies can provide potential benefits for a lot of applications such as text classification and clustering (Bloehdorn and Hotho, 2004), where additional conceptual features extracted from ontologies are used to enhance the bag-of-words model. In information retrieval and extraction, ontologies can solve the problem of vocabulary mismatch between documents and user queries, and many other problems (Guarino et al., 1999; Elabd et al., 2015). Also, ontologies provide the necessary concepts that represent a specific domain and state which semantic relations potentially hold between them (Nirenburg and Raskin, 2004).

For information systems that rely on Quranic Arabic, Quranic ontologies can be exploited very efficiently in improving the performance of the information extraction process. One such system is the question answering system for Quran (Q&A system) (Abdelnasser et al., 2014; Hakkoum and Raghay, 2016) where the user enters a question in Arabic related to Quran and the system retrieves the most relevant answer from Quran. Conventional Q&A systems are based on simple keyword search to find the best answer for the user query without knowing the meaning of the words and expressions used and the relationship between them. This kind of search has many limitations especially when applied to the Quranic Arabic where for example polysemy and synonymy are very common. Therefore, using ontologies to present the Quran knowledge provides semantics, logic, inference, and deep reasoning to the Q&A systems so complex and ambiguous queries could be solved very accurately.

In this paper, we introduce a novel approach that aims at enriching the automatic construction of Quran ontology. We extract the whole relations that exist in the conjunctive phrases which can be defined as any two words combined by the conjunction AND. The two words could be nouns, proper nouns, or adjectives. The main contribution is that we define a hybrid method to extract semantic relations from Quran based on strong and solid rules. First, we exploit an efficient rule in Arabic grammar, which is AND conjunction, to extract several types of semantic relations. AND conjunction is a well-known grammatical tool that combines terms which have a degree of association between each other. The proposed set of patterns is used to extract the AND conjunctive phrases from the Quranic Arabic Corpus and not to extract a specific type of semantic relations, as is known in the pattern-based methods. Second, we use an accurate measurement, namely the correlation coefficient, to find the association value between the combined words. This is totally new and useful in this field and different from other common measurements such as Mutual Information (MI) and *t*-score. Finally, we combine statistical tests (testing hypothesis and student *t*-test) and domain experts to validate the results achieved. All the reported approaches in the field of Quran mining (Alrehaali and Atwell, 2014) depend on either domain experts or exegesis books such as Tafsir of Ibn Kathir in the validation process. This step is very essential and cannot be neglected because the holy Quran is a very sensitive and critical text. Basically, we present a scientific validation approach to consolidate the domain experts' decisions.

Manually, we reveal three different categories of semantic relations from the entire Quran based on one type of Arabic patterns which is the conjunctive patterns. The rationale behind selecting conjunctives in general and the conjunction AND in particular stems from its importance in Arabic grammar. It is considered as the basic Arabic conjunctive for its frequent use and probable indications for different meanings in the sentence (Al-Taweel, 2009). We initiate the proposed work by converting the Quranic Arabic Corpus from Buckwalter transliteration scheme to Arabic script.

Then, we extract the set of words that form ontology terms. We cover nouns, proper nouns, and adjectives. Next, we apply a set of conjunctive patterns, which we previously defined, to extract the candidate relations from the corpus. A filtering approach based on a correlation coefficient is used to select strong relationships. Finally, we validate the proposed approach by using testing hypothesis and Student *t*-test besides domain experts. This statistical technique supports the output of domain experts in order to give the most accurate and correct results.

The rest of the paper is organized as follows. Section 2 presents related work. Preprocessing phase is described in Section 3 and ontology learning from the holy Quran phase is introduced in Section 4. Section 5 presents the validation phase. Experimental results are discussed in Section 6 and conclusion is presented in Section 7.

2. Related work

Ontology learning from text in general occupies a large area in computer science, whereas ontology learning from the holy Quran in particular suffers from the specific nature of the Arabic script and the depth of knowledge needed in this field (Habash, 2010). However, few recent Quranic studies were interested in developing approaches which accomplish ontology learning tasks and represent the Quranic knowledge in a semantic way as sets of concepts and relations. Duckes initiated the Quranic Arabic Corpus (QAC) which is the first online collaboratively constructed linguistic resource with multiple layers of annotation including part-of-speech tagging, morphological segmentation and syntactic analysis using dependency grammar (Dukes and Habash, 2010; Dukes and Atwell, 2012). Also, the author built ontology from (QAC) which finds relations between proper nouns or any nouns if they represent well-defined concepts such as the names of animals, locations, and religious entities. The ontology was validated based on scholarly sources, namely Tafsir of Ibn Kathir (Ibn Kathir, 1999).

A large corpus named QurAna (Sharaf and Atwell, 2012) was created from the original Quranic text, where specific types of words are considered as ontological concepts. Personal pronouns are extracted and tagged with their antecedents. These antecedents are maintained as an ontological list of concepts which improves information systems performance. Abbas exploited an existing index of Quranic topics from a scholarly source: Tafsir of Ibn Kathir, to develop Qurani (Abbas, 2009), which is a tool looks for concepts in the holy Quran and provides English translations for the verses containing these concepts. Yauri et al. proposed a system that reused Leeds ontology (Quranic Arabic Corpus ontology) to model Quran domain knowledge, using Web Ontology Language OWL (Yauri et al., 2012). The system added the act concepts related to specific topics in Quran such as praying, Zakat, sin and rewards, and showed the relations between them using Description Logic. The user of this model can semantically retrieve important concepts from the holy Quran. Verses referring to particular concepts could also be retrieved. Ul Ain and Basharat introduced DataQuest (Ul Ain and Basharat, 2011), an efficient framework for modelling and retrieving knowledge from distributed knowledge sources primarily related to the holy Quran and scholarly texts, with the use of semantic web, information extraction, and natural language processing techniques. The documents are annotated using the domain ontology. Thus, users can query that filtered and concise knowledge using a semantically based intelligent search engine. Another work which covers a specific topic in the holy Quran was conducted by Al-Yahya and colleagues to build a computational model for representing Arabic lexicons using ontologies (Al-Yahya et al., 2011). The model has been implemented on the Arabic language vocabulary associated with "Time

Download English Version:

<https://daneshyari.com/en/article/6899046>

Download Persian Version:

<https://daneshyari.com/article/6899046>

[Daneshyari.com](https://daneshyari.com)