Proceedings Paper:

![White Rose university consortium logo - Universities of Leeds, Sheffield & York]

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Computational ontologies for semantic tagging of the Quran:
A survey of past approaches.

Sameer M. Alrehaili, Eric Atwell

School of Computing
University of Leeds, Leeds LS2 9JT, UK
E-mail: salrehaili@gmail.com, e.s.atwell@leeds.ac.uk

Abstract

Recent advances in Text Mining and Natural Language Processing have enabled the development of semantic analysis for religious text, available online for free. The availability of information is a key factor in knowledge acquisition. Sharing information is an important reason for developing an ontology. This paper reports on a survey of recent Qur'an ontology research projects, comparing them in 9 criteria. We conclude that most of the ontologies built for the Qur'an are incomplete and/or focused in a limited specific domain. There is no clear consensus on the semantic annotation format, technology to be used, or how to verify or validate the results.

Keywords: information extraction, semantics, ontologies, Qur'an

1. Introduction

Recent advances in Text Mining and Natural Language Processing have led to a number of annotations for religious text such as the Qur'an.

Ontology-based models of computational semantics are being widely adopted in various fields such as Knowledge Management, Information Extraction, and the Semantic Web. Religious Studies researchers are also starting to exploit the ontology for improving the capture of knowledge from religious texts such as the Qur'an and Hadith. A definition of ontology in Artificial Intelligence is "the specification of conceptualizations, used to help programs and humans share knowledge.". Ontology development is generally described as an iterative process, and the development process never completes (Ullah Khan et al. 2013). Therefore, many researchers start with a focus on one or two semantic fields of their Qur'an ontology.

There are different annotations and ontologies already available for the Qur'an online and most of them are free. However, they differ in the format that they provide for End-Users, and in the technologies that they use to construct and implement the ontology. This variety of formats used to store the annotated data of the Qur'an leads to a gap between computer scientists, who make tools to provide analysis, and End-Users who are interested in the specific domain. Not all End-Users or linguistics researchers are technically able or willing to make their own converter. Therefore, the need to design a standard format and provide available analyses for the Qur'an in a standard format is becoming essential to facilitate End-User work and make them focus on the analysis instead of doing extra data-reformatting work. Moreover, this would increase the process of development in Qur'an analysis. This paper does not try to solve these challenges, instead of that it tries to survey the Qur'an ontology research projects that have been done recently, comparing them in terms of 9 criteria.

The rest of this paper is organized as follows: section 2 is a brief introduction about the Qur'an. Section 3 identifies the criteria for evaluation of Quran ontologies used in this survey. Section 4 discusses previous work related to the Qur'an and ontology. Finally, conclusion and research future directions are presented. This paper includes a comprehensive comparison table summarising the key features of the different existing ontologies of the Qur'an.

2. The Quran

Muslims believe that the Qur'an is God's word and the most widely read book in the world since its revelation; every Muslim can memorise and recite some parts of the Quran at least 17 times every day when praying. Its recitation and reading have not stopped one day since its revelation. The Qur'an includes a range of knowledge in different subjects such as science, art, stories and history, agriculture and industry, human and social relations, organization of finance, education and health. For some Muslims who do not speak Arabic, and for non-Muslims, the Qur'an is difficult to understand, although it has been translated into over 100 different languages.

3. Comparison criteria

Table 1 summarises the comparison of the Quran ontologies described in the literature. The comparison focuses on the content of ontologies in the work reviewed. The list of criteria used for comparison is described briefly in the following.

1. **Qur'an text:** The ontology relies on one of the following languages:
   - Original Arabic text (A).
   - English translation (B).
   - Malay translation (C).

This criterion has been chosen because we noticed that there is a variation of the language used in ontology-based work. For example (Ullah Khan et al. 2013) and (Saad et al. 2010) ontologies used English translations of the Qur'an, while (Ali & Ahmad 2013) used a Malay translation. This aspect should not be ignored in research

on reusing an ontology as it identifies a challenge in merging different translations of Qur'an ontologies.

2. **Coverage area:** Topics and word types that are covered by the ontology. For example, an ontology may covers the topic of animals for only nouns. This aspect compares the ontologies on the topic that they have created for.

3. **Coverage proportion:** This criterion identifies if the ontology covers the entire Qur'an or only some parts.
   - Entire the Qur'an (A).
   - Some parts (B).
   - Specific topic (C).

We found only one work that covers all Qur'an chapters, others focused on one or two topics.

4. **Underlying format:** There are many formats such as plain text, XML files, and RDF or OWL.

Format is also an important factor in ontology reuse due to requirements for extra work in extraction of ontology elements from the existing ontologies.

5. **Underlying technology used:** tools used for building and representing the ontology.

6. **Availability:** this criterion identifies if the ontology is free access or not. This is important in reusing an ontology too because we have noticed that there are some resources which are not available for download and reuse.
   - Yes (A)
   - No (B)

7. **Concepts number:** The number of abstract and concrete concepts in the ontology.

8. **Relations type:** The ontology may be have one of the following relations between the concepts:
   - Meronymy (Part-of) (A)
   - Synonyms (similar) (B)
   - Antonymy (opposite) (C)
   - Hyponymy (subordinate) (D)

9. **Verification method used:** The evaluation method used to verify the ontology. Two types of methods have been used to verify ontology-based work on the Qur'an:
   - Domain experts
   - Scholarly sources (Ibn Kathir)

This gives us information about quality of the work that has been conducted in order to evaluate the ontology.

4. Ontology research on the Qur'an

Several initial studies have been undertaken on the topic of Qur'an ontology. Most of these studies have been developed in order to improve the efficiency of information retrieval for the Qur'an. These have facilitated the process of accessing Qur'an knowledge. However, they vary from each other in different aspects such as coverage of the Qur'an, discourse level, language of the text used; original Arabic text or other translation, domain focused on, number of concepts and types covered, concept extraction method, relation types they provide, development process they followed during construction, technology used in ontology construction, availability, and verification method.

(Saad et al. 2009) proposed a simple methodology for automatic extraction of a concept based on the Qur'an in order to build an ontology. This paper used a method based on extracting the verses which contain a word of prayer in it as well as the previous and next verse. This method relies on a format of one English translation of the Qur'an that included some aspects such as Uppercase Letter. An uppercase letter is used to identify the concepts such as the Book. Another feature called compound noun is used to identify the relationship of hyponym or "Part-OF" between the concepts. A copula is used to identify the syntactic relationship between subject and adjective or noun. The ontology is based on the information obtained from domain experts. The development process is adopted from (Saad & Salim 2008). However, the authors have focused on the subject of Prayer or "Salat" particularly in daily prayer, thus this ontology does not cover all subjects in the Qur'an. In addition, there is no mention about underlying format or ontology technologies used in this paper.

Saad et al. continued their work to develop a framework for automated generation of Islamic knowledge concrete concepts that exist in the holy Qur'an as presented in (Saad et al. 2010). The framework takes into account some situations form the sciences of the Qur'an, such as the cause of revelation (Asbab Al Nuzul), and verses overridden by related verses that were revealed later (Nasikh Mansukh). The methodology of ontology development was also adopted from (Saad & Salim 2008), and the method to obtain the concepts is applying a grammar and extraction rules to the English translation of the Qur'an. The 374 extracted instances only cover verses that have the keyword salah or pray and this does not cover the entire Qur'an. These instances were mapped to six abstract concepts. This paper differs from the previous in synonym relations.

(Saad et al. 2011) proposed methods for designing an ontology based on translated texts of the Qur'an. Information used in developing the ontology was collected by the domain experts. Their ontology also only covers the subject of "Salat" (pray).

Another ongoing research project on a prototype of a framework called SemQ is presented in (Ai-khalifa & Ai-yahya 2009). SemQ identifies opposition relationships between Quranic concepts. The idea is SemQ receives a

verse as input and produces a list of words that are opposed to each other with the degree of the opposition. The coverage is in the domain of "Women" in the Qur'an. Ontology development makes use of the Buckwalter morphology POS annotation and focuses on nouns and verbs that are related to the semantic field of Time. This paper used OWL and UPON technologies in order to represent the concepts and relations. The ontology consists of seven abstract concepts and eleven concrete concepts. This ontology is sharable and can be downloaded. This study was limited to word level which includes only nouns and verbs of the "Women" domain. However, there are no evaluable results provided by the authors or any validation attempts for their proposed framework.

In (Ali & Ahmad 2013) , a theme-based approach is proposed to represent and classify the knowledge of the Qur'an using an ontology. Their ontology was developed according to themes described in Syammil Al-Quran Miracle the Reference, and using protégé-OWL and Malay language as medium of concepts, and was validated by the domain experts. It only covers two themes: "Iman" which means faith and "Akhlaq" which means deed. This was an Ontology-based approach to represent and classify Qur'anic concepts according to specific semantic fields. The structure of the ontology was verified by Qur'an domain experts. The ontology was developed using Protégé-OWL and using Malay Language as the medium language. The authors proposed a representation approach whcih differs from traditional representation which consist of Juz, Chapter and Verse. There is no explanation of what language was used for this ontology and what source the concepts were based on. They implemented the ontology using protégé. There are no details of results or validation of the ontology, although the paper states that the process of creating the ontology was reviewed by seven Qur'an domain experts.

 (Ullah Khan et al. 2013) developed a simple ontology for the Qur'an that includes the animals that are mentioned in the Qur'an in order to provide Qur'anic semantic search. The ontology was built using protégé editor, and SPARQL query language was used to retrieve the answers to a query. The English translation of the Qur'an by Pickthall is used in this ontology. The ontology provides 167 direct or indirect references to animals in the Qur'an obtained based on information mentioned in a book entitled "Hewanat-E-Qurani". The relationship type is a taxonomy relation. The paper concludes that the existing Arabic WordNet does not help for retrieving this type of document information.

(Yauri et al. 2012) proposed a model for defining the important Qur'anic concepts by knowledge representation and presented the relationships between them using Description Logic (Predicate logic). They reused the Quranic Arabic Corpus ontology by (Dukes 2013). This ongoing research attempts to reuse and improve an existing ontology developed in Leeds by adding more relations. Protégé is used in ontology construction. A top-down ontology development process was followed. It has 15 abstract concepts.

(Yauri et al. 2013) has proposed ontology-based semantic search for the Qur'an using protégée editor and Manchester OWL query language. The ontology was built by reusing the existing Quranic Arabic Corpus ontology developed by (Dukes 2013), and adding more than 650 relationships depending on the Qur'an, Hadith, and Islamic websites. This ontology was constructed manually.

(Yahya et al. 2013) proposed a semantic search for the Qur'an based on Cross Language Information Retrieval (CLIR). They created a bilingual ontology for the Quran composed of concepts based on existing Quranic Arabic Corpus ontology by (Dukes 2013), and found 5695 documents belonging to a main concept, where 541 documents are not assigned to any concepts in an English translation. In Malay, there are 5999 documents assigned to main concepts, where 237 documents do not belong to any concept.

In (Yunus et al. 2010), the authors did experiments on retrieving verses of the Quran using a semantic query approach exploiting Cross Language Information Retrieval (CLIR).

(Abbas 2009) developed a tool for searching for the Qur'anic concrete and abstract concepts. She exploited an existing Qur'an topics index from a scholarly source: Tafsir of Ibn Kathir. This onotology covered the whole Qur'an.

(Dukes 2013) in his PhD thesis defines 300 concepts in the Qur'an, and extracts the interrelationships using Predicate logic. The number of relations is 350. The type of relation between concepts is Part-of or IS-A. The ontology is also based on the Tafsir by Ibn Kathir.

(Muhammad 2012) in his thesis, developed an ontology covering the whole Qur'an in terms of pronoun tagging. Each pronoun is linked to its antecedent or previous subject.

In (Sharaf & Atwell 2012), the authors have created a dataset called QurSim which consists of 7600 pairs of related verses for evaluating the relatedness of short texts.

An automatic knowledge extraction method based on rules and natural language patterns is described in (Saad et al. 2013). Their methods rely on the English translation of the Qur'an and have identified a new pattern language named Qpattern which is suitable for extraction of taxonomy part-of relations. This research also identified that it is difficult to extract information from text that includes co-reference like the Qur'an.

The aim of this report was to look at the range of existing studies on Quran ontology available currently and identify the limitations of these studies as well as potential future work. Some semantic annotations have been done for the entire Qur'an, but for a specific type of word and domain, such as (Al-khalifa & Al-yahya 2009), an ontology for verbs in the domain of women, or the ontology of (Al-yahya et al. 2010) for nouns in the domain of time. There is one non-domain-specific ontology for the entire Qur'an but it is only for pronouns (Muhammad 2012). Most ontologies have relations using Part-Of or synonyms, but one work includes opposition relations, (Ai-khalifa & Ai-yahya 2009). Most ontologies built for the Qur'an are incomplete and focused in a specific domain. There is no clear consensus on the semantic annotation format, technology to be used, or how to verify or validate the results.

6.   References

Abbas, N., 2009. Quran'search for a Concept'Tool and Website. *MRes Thesis, School of Computing, Univeristy of Leeds.*

Al-khalifa, H. & Al-yahya, M.M., 2009. SemQ : A Proposed Framework for Representing Semantic Opposition in the Holy Quran using. *Current Trends in Information Technology (CTIT), 2009 International Conference on the*, pp.0–3.

Ali, B. & Ahmad, M., 2013. AL-QURAN THEMES CLASSIFICATION USING ONTOLOGY. *icoci.cms.net.my*, (074), pp.383–389. Available at: http://www.icoci.cms.net.my/proceedings/2013/PDF/P ID74.pdf [Accessed November 26, 2013].

Dukes, K., 2013. *Statistical Parsing by Machine Learning from a Classical Arabic Treebank*, PhD Thesis, School of Co mputing, University of Leeds

Muhammad, A.B., 2012. *Annotation of Conceptual Co-reference and Text Mining the Qur'an*, PhD Thesis, School of Computing, University of Leeds.

Saad, S. et al., 2010. A framework for Islamic knowledge via ontology representation. *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, pp.310–314. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?a rnumber=5466897.

Saad, S. & Salim, N., 2008. Methodology of Ontology Extraction for Islamic Knowledge Text. In *Postgraduate Annual Research Seminar, UTM*.

Saad, S., Salim, N. & Zainal, H., 2009. Pattern extraction for Islamic concept. *Electrical Engineering and …*, (August), pp.333–337. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5 254719 [Accessed November 26, 2013].

Saad, S., Salim, N. & Zainal, H., 2013. Rules and Natural Language Pattern in Extracting Quranic Knowledge. In *In the proceedings of the Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences (NOORIC 2013)*. Madinah, Sadudi Arabia: IT Research Center for the Holy Quran and Its Sciences (NOOR).

Saad, S., Salim, N. & Zainuddin, S., 2011. An early stage of knowledge acquisition based on Quranic text. *Semantic Technology and …*, (June), pp.130–136. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5 995777 [Accessed November 26, 2013].

Sharaf, A.-B. & Atwell, E., 2012. QurSim: A corpus for evaluation of relatedness in short texts. In N. C. (Conference Chair) et al., eds. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Ullah Khan, H. et al., 2013. Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*, 2(6), pp.570–575. Available at: http://www.ijfcc.org/index.php?m=content&c=index& a=show&catid=43&id=493 [Accessed December 16, 2013].

Yahya, Z. et al., 2013. Query Translation Using Concepts Similarity Based on Quran Ontology for Cross-Language Information Retrieval. *Journal of Computer Science*, 9(7), pp.889–897. Available at: http://thescipub.com/abstract/10.3844/jcssp.2013.889. 897 [Accessed November 7, 2013].

Yauri, A.R. et al., 2013. Ontology Semantic Approach to Extraction of knowledge from Holy Quran. *Computer Science and Information Technology (CSIT), 2013 5th International Conference*, pp.19–23. Available at: http://0-ieeexplore.ieee.org.wam.leeds.ac.uk/stamp/sta mp.jsp?tp=&arnumber=6588752&isnumber=6588741.

Yauri, A.R. et al., 2012. Quranic-based concepts: Verse relations extraction using Manchester OWL syntax. *2012 International Conference on Information Retrieval & Knowledge Management*, pp.317–321. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?a rnumber=6204998.

Yunus, M. a, Zainuddin, R. & Abdullah, N., 2010. Semantic query for Quran documents results. In *2010 IEEE Conference on Open Systems (ICOS 2010)*. Ieee, pp. 5–7. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?a rnumber=5719959.

Reference	1	2	3	4	5	6	7	8	9
(Saad et al. 2009)	B	"Pray"	C	N/A	N/A	N/A	N/A	A, Part-Of	Domain experts
(Saad et al. 2010)	B	"Pray"	C	N/A	N/A	N/A	374 instances and 6 abstract	B, synonyms	Domain experts
(Ai-khalifa & Ai-yahya 2009)	A	"Women", Nouns and Verb	C	OWL	UPON	A	11 concrete and 7 abstract	C, opposition	N/A
(Al-yahya et al. 2010)	A	"Time noun"	C	OWL	UPON	A	11 concrete and 7 abstract	D, hyponymy	N/A
(Ali & Ahmad 2013)	C	"faith and deed"	C	OWL	protégé	N/A	N/A	N/A	Domain experts
(Ullah Khan et al. 2013)	B	"animals"	C		Protégé, SPARQL	N/A	N/A	A, Part-od	N/A
(Yauri et al. 2012)	N/A	"salat, zakat, sin, reward"	C	OWL	Protégé	N/A	15 abstract	N/A	N/A
(Yauri et al. 2013)	N/A	N/A	N/A	Manchester OWL	N/A	N/A	N/A	N/A	Manually constructed
(Yahya et al. 2013)	B, C	N/A	C	N/A	N/A	N/A	N/A	N/A	N/A
(Yunus et al. 2010)	A, B, C	N/A	C	N/A	N/A	N/A	N/A	N/A	N/A
(Dukes 2013)	A, B	N/A	N/A	Text files	N/A	B	300	Part-of	Ibn Kathir
(Muhammad 2012)	A	pronouns	A	XML	N/A	A	N/A	N/A	Ibn Kathir

Table 1: summary of ontology features in papers reviewed