Proceedings Paper:
Alqahtani, M and Atwell, ES orcid.org/0000-0001-9395-3764 (2016) Arabic Quranic Search Tool Based on Ontology. In: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016 Proceedings, Natural Language Processing and Information Systems, Lecture Notes in Computer Science. NLDB 2016 21st International Conference on Applications of Natural Language to Information Systems,, 22 Jun 2016, Salford, UK. Springer International Publishing , pp. 478-485. ISBN 978-3-319-41753-0

https://doi.org/10.1007/978-3-319-41754-7_52

Arabic Quranic Search Tool Based On Ontology

Mohammad Alqahtani, Eric Atwell

School of Computing, University of Leeds, Leeds, UK
{scmmal, E.S.Atwell} @leeds.ac.uk

Abstract. This paper reviews and classifies most of the common types of search techniques that have been applied on the Holy Quran. Then, it addresses the limitations of these methods. Additionally, this paper surveys most existing Quranic ontologies and what are their deficiencies. Finally, it explains a new search tool called: a semantic search tool for Al-Quran based on Qur'anic ontologies. This tool will overcome all limitations in the existing Quranic search applications.

Keywords: Holy Quran; Information Retrieval (IR); Natural Language Processing (NLP); Ontology; Semantic Search

1 Introduction

The Holy Quran (Al Quran) is sacred Arabic text [1]. Al Quran contains about 79,000 words forming 114 chapters [1]. The techniques used to retrieve information from Al Quran can be classified into two types: a semantic-based and a keyword-based technique. The semantic-based technique is a concept-based search tool that retrieves results based on word meaning, or concept match, whereas the keyword-based technique returns results based on letters matching word(s) queries [2]. The majority of Quranic search tools employ the keyword search technique.

The existing Quranic semantic search techniques are: an ontology-based [3], a synonyms-set [4] and a cross-language information retrieval (CLIR) technique [5]. The ontology-based approach searches for the concept(s) matching a user words query. Then, this technique returns verses related to these concept(s). The synonyms-set technique produces all synonyms of the query words using WordNet. After that, it finds all Quranic verses matching these words' synonyms. CLIR translates words of an input query to another language and then retrieves verses that contain words matching the translated words.

Several deficiencies exist with the Quranic verses (Ayat) retrieved for a query using the existing keyword search techniques. These problems are: irrelevant verses are retrieved, relevant verses are not retrieved or the order of retrieved verses is not ranked [4]. The keyword-based technique's limitations include misunderstanding the exact meaning of input words forming a query and neglecting some theories of information retrieval [8].

Moreover, current Quranic semantic search techniques have some limitations about finding requested information. This is because these semantic searches use uncom-

adfa, p. 1, 2011.

pleted Holy Quran ontology. Additionally, these concepts have different scopes and formats [9].

This paper is organized as follows. Section 2 is Literature review containing a review of the structure of the Holy Quran, Quranic search applications, previous research on Quranic search tools and existing ontologies of Al Quran. Section 3 describes the methodology of Arabic Quranic Search Tool Based on Ontology. Finally, Section 4 concludes the critical points in this paper.

2 Literature review

2.1 Structure of the Holy Quran

Challenging points regarding the natural structure of the Holy Quran exist when applying NLP technologies. First, a concept could be mentioned in different verses. For example, the concept of the Hell (النـار) is discussed in various chapters and verses. Additionally, one verse may contain many themes. For example, verse 40 of Chapter 78 contains only seven words describing five different concepts such as Allah, Humans, chastisement, person and, the Judgment day [8]. Another unique style of the Quran is that one concept is mentioned using different words, depending on the context. For example, Muhammad (محمد) is the same as Ahmad (أحمد), and Mozzammil (مُزَّمّل). Additionally, a term may also refer to entirely different things, depending on the context: for example, L-jannat 'الْجَنَّة' might refer to a heaven or a garden. Additionally, two different words may have the same letters but have different diacritics. For example, 'الجنة' represents two distinct words: 'الْجَنَّة' means paradise, and 'الجِنَّة' means ghosts (see Table 1).

Table 1. Search results of الجنة based on concepts

Arabic Pronunciation	English meaning	Arabic word	No. of verses
L-jannat	Paradise	الجَنة	109
aṣḥābu L-jannat	Companions of Paradise	اصحاب الجنة	14
L-jannat	Garden	الجَنة	14
Jinnat	Ghost	الجِّنة	10
Junnat	cover	الجُّنة	2

Finally, the text of the Holy Quran is written in the classical Arabic language, which is slightly different from the modern Arabic language. This will cause a gap between the query and retrieved verses.

2.2 Quranic Search Applications

Desktop and Web applications have been developed to retrieve knowledge from Al Quran. The majority of these applications use keyword search techniques. However, some researchers have proposed frameworks for a Quranic semantic search tool based on concepts.

Khazain-ul-Hidayat[10], and Zakr [11] are free desktop applications that enable user to read, listen to and search the Quran in many different languages. These applications are mainly designed to be aid tools for teaching the Quran. A user can search Al Quran by querying a word or by entering a verse number. When the user queries a word, the results will include all verses containing any forms of this query word based on the stem of query words. For example, if the query word is 'ذكر', then the retrieved verses will contain other forms of 'ذكر', such as 'اذكر', 'تذكرة', 'تذكرة', 'الذاكرون', and 'ذكرى'.

Almonagib alqurany (المنقب القراني) [12], Islam web [13], Tanzil [14], Quranic Arabic Corpus (QAC) [15], KSU Quran [17], The Quran [18] and the Noble Quran [16] are online Web applications that enable users to read, listen to and search Al Quran in different languages. Users can select a specific chapter, verse or word. In the case of searching by a word, these applications will return all verses that have words belonging to the same root of the query word.

Semantic Quran [19] is an online search tool application that allows a user to search verses based on concepts. In this application, each verse has a set of tags that are concepts. Additionally, not all verses are completely tagged. Therefore, the user can participate in tagging any verse. The idea behind this application is that many verses in the Holy Quran relate to certain concepts even though these verses do not have any common words.

2.3 Conducted Research on Quranic Search Tools

A lot of computational research has been carried out on Al Quran. The following review is about the IR and semantic search research.

[20] proposed a new Arabic question-answering system in the domain of Al Quran. The system prompts users to enter an Arabic question about Al Quran. Then, this system retrieves relevant Quranic verses with their Arabic descriptions from Ibn Kathir's book. This system uses 1,217 Quranic concepts integrated from the Quranic Arabic Corpus Ontology [15] and Quranic Topic Ontology [21]. It is claimed that retrieved results' accuracy can reach 65% using the top result. This system has three phases for answering a question: question analysis using the 'Morphological Analysis and Disambiguation of Arabic' tool (MADA) [22], IR using 'explicit analysis approach' [23] and answer extraction. This proposed system does not recommend a solution for if the question terms do not match any concepts from the Quranic Ontology.

[24] suggested a Quranic semantic search tool by developing a simple domain of ontology for the animals mentioned in Al Quran. This paper concludes that the existing Arabic WordNet is not sufficient for finding synonyms for query words to increase one's chances of retrieving information from a document, and based on this, it is suggested to develop Arabic WordNet for the Al Quran words.

[25] proposed a semantic search system for retrieving Quranic verses based on the enhanced ontology done by [15].

[26] recommended a semantic search for Al Quran based on CLIR. This research created a bilingual ontology: the English and Malay languages. This ontology is

based on the ontology developed by [15]. They did this to experiment on this ontology for two translations of Al Quran. In the Malay translation, 5,999 verses are assigned to the concepts, and 237 verses do not relate to any concepts. In the English translation, they found 5,695 verses related to concepts in this ontology. On the other hand, 541 documents are not allocated to any concepts.

[27] has developed a tool called "Qurany" for searching the Quranic text in both Arabic and English. In this project, 6,236 HTML pages were created in which each HTML page contains one verse, its eight different English translations and the topic of this verse in both Arabic and English. The project's main idea is searching Al Quran's eight translations using the keyword search. This will enhance this tool's precision. Regarding [27], most of the available search tools on the Web use one English translation in the search process, with average recall and precision values of 54% and 48%, respectively.

In conclusion, all of the above research on Quranic semantic search are proposed frameworks for developing a semantic tool to search Al Quran. Moreover, no online tools for Quranic semantic search are currently up-to-date.

2.4 Research on Ontology of The Holy Quran

[28] developed a Semantic Quran dataset in an RDF format representing 42 different Al Quran translations. This dataset was built by merging data from two different semi-structured sources: the Tanzil project and the Quranic Arabic Corpus. This ontology has 7,718 links to DBpedia, 18,655 links to Wiktionary and 15,741,399 triples.

[24] developed an ontology for Al Quran in the scope of the animals found in Al Quran. The ontology provides 167 links to animals in Al Quran based on information found in "the Hewanat Al-Quran" book (حيوانات القران).

[29] rebuilt the existing ontology created by [15] using the Protégée tool and Manchester OWL. He increased the number of relationships from 350 to about 650 based on Al Quran, the Hadith and some online Islamic resource.

[27] has developed a tree of nearly 1,100 nodes representing Quranic abstract concepts. These concepts are linked to all verses of Al Quran. She used existing Quranic topics from the Islamic scholarly book called "Mushaf Al Tajweed" (مصحف التجويد). These concepts in the index have an aggregation relationship; the hierarchy of concepts is non-reflexive, non-symmetric and transitive.

[15] extracted 300 concepts and 350 relations from Al Quran using predicate logic. The relationship types connecting concepts are Part-of and IS-A. The ontology is based on a famous Al Quran discerption book called "Tafsir Ibn Kathir" (تفسير ابن كثير) [30].

[31] developed an ontology for Al Quran in the scope of pronoun antecedents. This ontology consists of 1,050 concepts and more than 2,700 relations. Additionally, the relationship types connecting concepts are has-antecedent, has-concept and has–a-segment. Additionally, he produced a dataset called "QurSim" containing 7,600 pairs of related verses that have similarity in the main topic. The scope of this dataset is the similarity of verses.

[32] unified Al Quran Arabic Corpus [15], Quran annotated with Pronominal Anaphora [30] and the Qurany project [27]. These datasets are merged in one XML file, and then the file is uploaded in the Sketch Engine tool as a unified Quranic corpus.

In conclusion, all developed Quranic ontologies have different scopes, such as animals, verses' similarity and Quranic topics. Additionally, these ontologies were built in different formats, such as XML, RDF and OWL. Moreover, these various datasets have some similarity in concepts. To overcome the limitations in existing AL-Quran Ontologies, we aligned the four main ontologies [15], [27], [28] and [31]. Before the process of matching Ontologies, all ontologies is normalized to one file format. After unifying ontologies in one file format, the process of alignment follows the methodology of aligning ontology in [36] and [37].

3 Arabic Quranic semantic search tool

Figure1 is a framework for a new semantic search tool called Arabic Quranic Semantic Search Tool based on ontology (AQSST). This search tool aims to employ both IR techniques and semantic search technologies. The design of this tool is constructed based on the theories in proposed previous research [25], [33]–[35]. AQSST is divided into six components: Quranic Ontology (QO), Quranic Database QDB, Natural Language Analyser (NLA), Semantic Search Model (SSM), Keyword Search Model (KSM) and Scoring and Ranking Model (SRM).

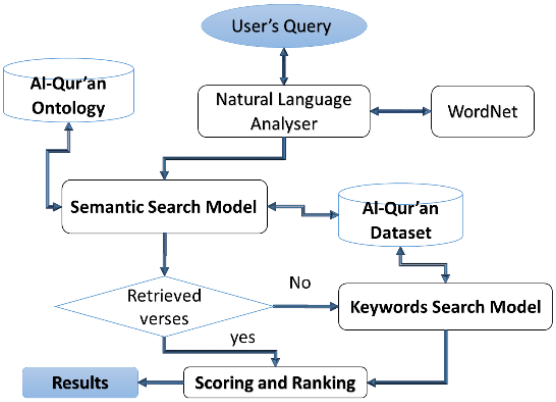

Fig. 1. Arabic Quranic Semantic Search tool structure (AQSST)

Quranic Ontology (QO) contains the new aligned Quranic ontology. QDB consists of Al Quran text in Arabic language and eight English translations of Al Quran, four different Tafsir (Description of AL-Quran), Al Quran words dictionary, Revelation reasons, concepts in Al Quran, and Named Entities (NE) based on Al Quran domain.

A user query undergoes different processes in NLA Model. This NLA parses a Natural Language query and then applies different NLP techniques on the tokenized

query. These techniques are: spell correction, stop word removal, stemming and Part Of Speech (POS) tanging. After that, NLA uses Arabic WordNet to generate synonyms for the reformatted query words. Then, NLA adds semantic tags to these words using NE list, as shown in table 2, and then sends the results to SSM.

Table 2. Example of analysing words in a query

Word	POS	Stem	Semantic Tag	Synonym	Weight
الرحمن	NOUN	رحمن	Allah Name		2
الذئب	NOUN	ذئب	Animal	السرحان	2
يوسف	NOUN	يوسف	Prophet		2
البئر	NOUN	بئر	Well	الجُبّ	2
محمد	NOUN	محمد	Prophet	احمد	2

SSM searches the Quranic ontology dataset by using SPARQL to find concepts related to the normalised query and then returns result to SRM. However, if no result is found KSM searches for verses contains words matching the analysed input words.

SRM filters the retrieved results from SSM and KSM; by eliminating the redundant verses (aya'at). Next, SRM ranks and scores the refined results based on the number of matching words in the results, the NE type of both the question and the answer, and the short distance between matched expressions in the retrieved results and question words. Finally, SRM provides the results to the user and then records the selected result. For instance, if a user search for l-jannat "الجنة" the retrieved result as shown in figure 2.

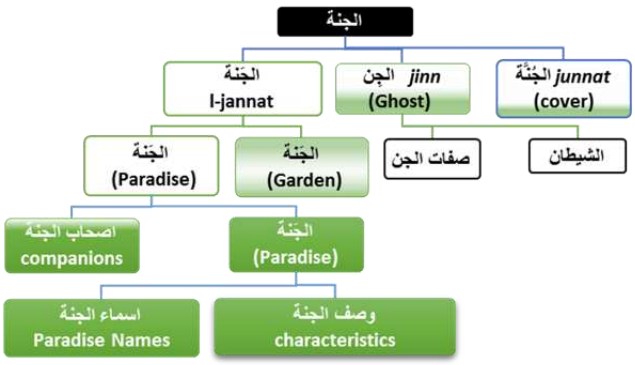

Fig. 2. Search result of (الجنة) in AQSST (dark colour is the most relevant result)

4 Conclusion

This paper summarises the search techniques used in existing search tools for AL Quran. Additionally, this paper studied the previous research have been conducted on Al Quran search methods and Al Ontologies. Depending on this study, many challenges have been found.

Firstly, Limitations of existing Al Quran search tools for retrieving all requested information. These search tools do not prompt users to search by concepts, phrases, sentences, questions or topics. Most search tools do not analyse the query texts by applying NLP techniques, such as parsing and spell check.

Secondly, absence of accurate and comprehensive resources for Islamic ontology. Existing Quranic datasets have different scopes and formats; and not follow ontology standards. Additionally, some Quranic ontologies are not available for usage.

The NER of the Arabic language is mostly focused on the modern Arabic language. Additionally, no well-formatted NER lists exist that are specialised for Quranic text, such as prophets' names, Allah's names, animals, times, religion, and etc. The Arabic language is an inflected language with complicated orthography.

Finally, all these limitations are highly considered in the design of the Arabic Semantic Quranic search tools.

References

1. E. Atwell, C. Brierley, K. Dukes, M. Sawalha, and A.-B. Sharaf, "An Artificial Intelligence Approach to Arabic and Islamic content on the Internet," in Proceedings of NITS 3rd National Information Technology Symposium, 2011.
2. G. Sudeepthi, G. Anuradha, and M. S. P. Babu, "A Survey on Semantic Web Search Engine," Int. J. Comput. Sci., vol. 9, 2012.
3. A. R. Yauri, R. A. Kadir, A. Azman, and M. A. A. Murad, "Quranic Verse Extraction base on Concepts using OWL-DL Ontology," Res. J. Appl. Sci. Eng. Technol., vol. 6, no. 23, pp. 4492–4498, 2013.
4. M. Shoaib, M. Nadeem Yasin, U. K. Hikmat, M. I. Saeed, and M. S. H. Khiyal, "Relational WordNet model for semantic search in Holy Quran," in Emerging Technologies, 2009. ICET 2009. International Conference on, 2009, pp. 29–34.
5. M. A. Yunus, R. Zainuddin, and N. Abdullah, "Semantic query for Quran documents results," in Open Systems (ICOS), 2010 IEEE Conference on, 2010, pp. 1–5.
6. A. Al Gharaibeh, A. Al Taani, and I. Alsmadi, "The Usage of Formal Methods in Quran Search System," in Proceedings of International Conference on Information and Communication Systems, Ibrid, Jordan, 2011, pp. 22–24.
7. B. Abu Shawar and E. Atwell, "An Arabic chatbot giving answers from the Qur'an," Proc. TALN04, vol. 2, pp. 197–202, 2004.
8. S. A. Raza, M. Rehan, A. Farooq, S. M. Ahsan, and M. S. Khan, "AN ESSENTIAL FRAMEWORK FOR CONCEPT BASED EVOLUTIONARY QURANIC SEARCH ENGINE (CEQSE)," 2014.
9. S. M. Alrehaili and E. Atwell, "Computational ontologies for semantic tagging of the Quran: A survey of past approaches," Lr. 2014 Proc., 2014.
10. "Khazain-ul-Hidayat." [Online]. Available: http://www.khazainulhidayat.com/.
11. "Zakr." [Online]. Available: http://zekr.org/.
12. "Almonagib alqurany." [Online]. Available: http://www.holyquran.net/search/sindex.php.
13. "Islam web." [Online]. Available: http://quran.al-islam.com/.
14. H. Zarrabi-Zadeh, "Tanzil," 2007. [Online]. Available: http://tanzil.net/.
15. K. Dukes, "Statistical Parsing by Machine Learning from a Classical Arabic Treebank," School of Computing, University of Leeds, 2013.

16. “Noble Quran.” [Online]. Available: http://quran.com/.
17. “KSU Quran.” [Online]. Available: http://quran.ksu.edu.sa/.
18. “The Quran.” [Online]. Available: http://al-quran.info/.
19. “Semantic Quran.” [Online]. Available: http://semquran.com/.
20. H. Abdelnasser, R. Mohamed, M. Ragab, A. Mohamed, B. Farouk, N. El-Makky, and M. Torki, “Al-Bayan: An Arabic Question Answering System for the Holy Quran,” ANLP 2014, p. 57, 2014.
21. N. H. Abbas, “Quran ‘search for a concept’ tool and website,” 2009.
22. N. Habash, O. Rambow, and R. Roth, “MADA+ TOKAN: A toolkit for Arabic Tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization,” in Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, 2009, pp. 102–109.
23. E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis,” in IJCAI, 2007, vol. 7, pp. 1606–1611.
24. H. U. Khan, S. M. Saqlain, M. Shoaib, and M. Sher, “Ontology Based Semantic Search in Holy Quran,” Int. J. Futur. Comput. Commun., vol. 2, no. 6, pp. 570–575, 2013.
25. A. R. Yauri, “Automated Semantic Query Formulation for Quranic Verse Retrieval,” Putra Malaysia, Malaysia, 2014.
26. Z. Yahya, M. T. Abdullah, A. Azman, and R. A. Kadir, “Query Translation Using Concepts Similarity Based on Quran Ontology for Cross-Language Information Retrieval,” J. Comput. Sci., vol. 9, no. 7, pp. 889–897, Jul. 2013.
27. N. Abbas, “Quran’search for a Concept'Tool and Website,” Unpubl. Diss. Leed, 2009.
28. M. A. Sherif and A.-C. Ngonga Ngomo, “Semantic QuranA multilingual resource for natural-language processing,” Semant. Web, 2009.
29. A. R. Yauri, R. A. Kadir, A. Azman, and M. A. A. Murad, “Ontology semantic approach to extraction of knowledge from Holy Quran,” in Computer Science and Information Technology (CSIT), 2013 5th International Conference on, 2013, pp. 19–23.
30. M. S. Abdul-Rahman, Tafsir Ibn Kathir Juz’ 1 (Part 1): Al-Fatihah 1 to Al-Baqarah 141 2nd Edition. MSA Publication Limited, 2009.
31. A.-B. M. Sharaf and E. Atwell, “QurAna: Corpus of the Quran annotated with Pronominal Anaphora,” in LREC, 2012, pp. 130–137.
32. L. Aldhubayi, “Unified Quranic Annotations and Ontologies,” University of Leeds, UK, 2012.
33. M. Mony, J. Rao, and M. Potey, “Semantic Search based on Ontology Alignment for Information Retrieval,” Int. J. …, 2014.
34. 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., vol. 35, no. 2, p. 21, 2010.
35. S. Ou, V. Pekar, C. Orasan, C. Spurk, and M. Negri, “Development and Alignment of a Domain-Specific Ontology for Question Answering.,” Lr. 2008.
36. A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy, “[GLUE] Learning to Match Ontologies on the Semantic Web,” VLDB J. Int. J. Very Large Data Bases, vol. 12, no. 4, pp. 303–319, 2003.
37. M. Taye and N. Alalwan, “Ontology Alignment Technique for Improving Semantic Integration,” … Conf. Adv. Semant. …, no. c, pp. 13–18, 2010.