# Quranic Arabic Semantic Search Model Based on Ontology of Concepts

Mohammad Mushabbab A. Alqahtani

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Computing

January 2019

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

*Assertion of moral rights (optional):*

The right of Mohammad Mushabbab A. Alqahtani to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

# Publications

Chapters 3,4,6 of this thesis are based on jointly-author publications. The candidate is the principal author of all original contributions presented in these papers, the co-authors acted in an advisory capacity, providing feedback, general guidance and comments.

**Chapter 3**

The work in chapter 3 of the thesis has appeared in publication as follows:

Alqahtani, M., & Atwell, E. (2017). Evaluation Criteria for Computational Quran Search. International Journal on Islamic Applications in Computer Science and Technology, 5(1). Retrieved from http://www.sign-ificance.co.uk/index.php/IJASAT/article/view/1562.

Alqahtani MMA; Atwell E (2016) Comparison Criteria for Computational Quranic Search Methods. In 4th International Conference on Islamic Applications in Computer Science and Technologies, Sudan, 20 Dec 2016 - 22 Dec 2016.

**Chapter 4**

The work in chapter 4 of the thesis has appeared in publication as follows:

Alqahtani MMA; Atwell E (2016) Aligning and Merging Ontology in the Quran Domain. In the 9th Saudi Students conference in the UK, Birmingham, 13 – 14 Feb 2016.

Alqahtani, M. M., & Atwell, E. (2018, March). Developing Bilingual Arabic-English Ontologies of the Quran. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)* (pp. 96-101). IEEE.

**Chapter 6**

The work in chapter 6 of the thesis has appeared in publications as follows:

Alqahtani, M.M.A., Atwell, E. (2015). A Review of Semantic Search Methods to Retrieve Information from the Qur’an Corpus. In 8th international Corpus Linguistics Conference, Lancaster, pp. 21–24, Jul 2015.

Alqahtani, M.M.A., Atwell, E. (2015). Quranic search tool based on the ontology of concepts. In the 8th Saudi Students Conference in London, pp. 29–30, Jan 2015.

Alqahtani, M., & Atwell, E. S. (2016, June). 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. Vol. 9612, pp. 478–485: Springer International.

# Acknowledgements

*First and foremost, I thank and praise Allah for providing me with the patience, strength, wellbeing, and skills to complete this thesis at one of the most important stages of my life.*

*Undertaking this thesis has been a truly life-changing experience, and it would not have been possible without the support and guidance that I received from a number of people.*

*First, I would like to thank my supervisor, Professor Eric Atwell, for his support over the last four years. Without his guidance and valuable feedback, completing this thesis would not have been achievable. Professor Atwell, thank you for your continuous encouragement, which has allowed me to publish most of my original contributions.*

*My deepest gratitude goes also to my mother and to my wife, Rawia, to whom this thesis is dedicated, for everything they have provided, and for their love, patience, and unfailing support.*

*My sincere gratitude and thanks are also directed toward colleagues, friends, Arabic NLP group members, and members of the research community, who gave me invaluable advice, encouragement, and support, and for the valuable seminars we enjoyed.*

*I would also like to acknowledge the University of Jeddah for granting me a scholarship to pursue my studies in the UK.*

# Abstract

The Holy Quran is the essential resource for Islamic sciences and Arabic language. Therefore, numerous Quranic search applications have been built to facilitate the retrieval of knowledge from the Quran. This thesis presents a novel Arabic Quran semantic search model.

First, this thesis evaluated existing search tools constructed for the Holy Quran, against 13 criteria depending on: search features, output features, the precision of the retrieved verses, recall database size, and types of database contents.

Then, the study reviewed the existing Quran ontologies and compared them against 11 criteria. Some deficits have been found in all these ontologies. Additionally, a single Quranic ontology does not cover most of the knowledge in the Quran. Therefore, I developed a new Arabic-English Quran ontology from ten datasets related to the Quran such as: Quran chapter and verse names, Quran word meanings, and Quran topics. The main aim of developing a Quranic ontology is to facilitate the retrieval of knowledge from the Quran. Additionally, the Quran ontology will enrich the raw Arabic and English Quran text with Islamic semantic tags.

Furthermore, I developed the first Annotated Corpus of Quran Questions and Answers in Arabic. This corpus has 2200 pairs of question and answer collected from trusted Islamic sources. Each pair of question and answer is labelled with 5 tags. Examples of tags are: question type: either factoid or descriptive, topic of question-based on the Quran ontology, and question class.

Finally, the thesis explains a new semantic search model for the Arabic Quran based on my Quran ontology. This model aims at overcoming limitations in the existing Quran search applications. This search tool employs both Information Retrieval techniques and semantic search technologies. The performance of this search model is evaluated by using The Annotated Corpus of Arabic Quran Questions and Answers.

# Table of Contents

## List of Tables

## List of Figures

## List of Abbreviations

The following table lists the meanings of the abbreviations used in this thesis.

| Abbreviation | Meaning |
|---|---|
| AAQQAC | Annotated Arabic Quran Question and Answer Corpus |
| AAQQAC | Annotated Arabic Quranic Question-Answer Corpus |
| ACL | Arabic Learner Corpus |
| AQC | Arabic Quran Corpus |
| AQQAC | Arabic Quranic Question-Answer Corpus |
| CE | Common Era |
| CLIR | Cross-Language Information Retrieval |
| CVS | Comma-Separated Values |
| DAML | Darpa Agent Markup Language |
| DBMS | Database Management Systems |
| DiDOn | Diagrams Into Domain Ontology |
| DL | Deep Learning |
| Fact++ | Fast Classification Of Terminologies |
| FAQ | Frequently-Asked Questions |
| IR | Information Retrieval |
| KSM | Keyword Search Model |
| MADA | Morphological Analysis and Disambiguation Of Arabic |
| MAP | Mean Average Precision |
| MOKI | Modelling Wiki |
| MSA | Modern Standard Arabic |
| NE | Named Entity |
| NeON | Network Of Ontologies |
| NER | Named Entity Recognition |
| NL | Natural Language |
| NLA | Natural Language Analyser |
| NLP | Natural Language Processing |
| OAQQCT | Online Arabic Quranic Question Classifier Tool |
| OIL | Ontology Inference Layer |
| OQC | Ontology Of Quranic Concepts |
| OWL | Web Ontology Language |
| PHP | Personal Home Page( Hypertext Pre-processor) |
| POS | Part Of Speech |

| | |
|---|---|
| Q&A | Question and Answer |
| QAC | Question & Answer Corpora |
| QCO | Quranic Arabic Corpus Ontology |
| QDB | Quran Database |
| QO | Quran Ontology |
| QT | Quranic Topics |
| QurAna | Quran Annotated With Pronominal Anaphor |
| QurSim | Quran Similarity |
| RACER | Renamed A Box and Concept Expression Reasoner |
| RDF | Resource Description Framework |
| RDFS | Resource Description Framework Schema |
| SPARQL | Simple Protocol and RDF Query Language |
| SRM | Scoring and Ranking Model |
| SSM | Semantic Search Model |
| SVM | Support Vector Machine |
| SW | Semantic Web |
| SWOOP | Semantic Web Ontology Overview and Perusal |
| TopQuadrant | Top Quadrant |
| TREC | Text Retrieval Conference |
| XML | Extensible Markup Language |

# Chapter 1
# Introduction

This chapter will summarise the research problems that are addressed in this thesis, with an explanation of the field of study. Moreover, it will elaborate on the motivations and rationale for choosing the Quran as the topic for this Ph.D. project. Then, the scope of this project will be presented, and the novelty and original contributions of this research highlighted. Finally, the structure of the thesis will be set out.

## 1.1 Background of the study

Information retrieval (IR) can be defined as methods for finding required information within large collections of text, usually stored on computers (Manning, Raghavan & Schütze, 2008). As early as 1945, Vannevar Bush suggested using computers to search for specific information (Singhal, 2001), and the first automated IR systems were developed in the 1950s. Most people engage in IR daily when they use World Wide Web search engines, such as Google or Yahoo. The main practical goals of IR are: to enable users to search huge collections of data quickly; to provide for more flexible matching algorithms, such as stemming query words; and to rank retrieved results.

IR systems use many techniques to increase the quality, precision, and recall of search results. Examples of these methods are inverted indexes, pre-processing before document indexing, sorting results, term weighting, the vector space model, and evaluation of an IR system based on the relevance of the documents it retrieves. However, ambiguity in the search results persists as a major issue in IR.

Semantic search is a newer version of IR, one which works by reasoning an input query applying a knowledge base and returning the most relevant answers. The input query can take different forms, such as a natural language question, a triple knowledge representation of a question, a graphical representation, or keywords. The knowledge base can involve one or more ontologies, corpora, or plain text documents. Similarly, the answers retrieved from a semantic search can take a multitude of forms, from pure triples to a natural language representation.

Both the IR and semantic search techniques have been applied to the Holy Quran. The Quran is the essential resource for the Islamic sciences and Arabic language. Muslims believe that the Quran is the product of a revelation from Allah over 1300 years ago, from 609 to 632 CE. It contains 77439 words over 114 chapters (Atwell, Brierley, Dukes, Sawalha & Sharaf, 2011). Each chapter consists of a varying number of verses. Additionally, the Quran contains 6236 verses (Ayat). The Quran has various topics, such as ethics, the law of Islam and marital and family laws. These features create a rich environment for a semantic search.

Many Quran search programs have been built to facilitate the retrieval of knowledge from the Quran. In previous studies, the techniques that were used to retrieve information from the Quran can be classified into two types: semantic-based, and keyword-based techniques. The semantic-based technique is a concept-based search tool that retrieves results based on word meaning, or concept match. By contrast, the keyword-based technique returns results based on letters matching word(s) queries (Sudeepthi, Anuradha, & Babu, 2012). The majority of Quran search tools employ the keyword search technique.

The existing Quran semantic search techniques consist of the ontology-based (Yauri, Kadir, Azman & Murad, 2013), synonyms-set (Shoaib, Nadeem Yasin, Hikmat, Saeed & Khiyal, 2009), and cross-language information retrieval (CLIR) techniques (Yunus, Zainuddin & Abdullah, 2010). The ontology-based technique searches for the concept(s) matching a user query and then returns the verses related to this concept(s). The synonyms-set method produces all synonyms of the queried word using WordNet and then finds all verses in the Quran matching these synonyms. The CLIR technique translates the words of an input query into a specified language and then retrieves the verses that contain words matching the translated words.

On the other hand, text-based techniques comprise the keyword-matching, morphological-based (Al Gharaibeh, Al Taani & Alsmadi, 2011), and chatbot techniques (Abu Shawar & Atwell, 2004). The keyword-matching method returns verses that contain any of the queried words. The morphological-based method consists of a root word search; this generates all other forms of the query word and then finds all verses in the Quran that match these word forms. The chatbot method selects the most significant or important words from a user query and then returns the Quranic verses containing any words matching the selected words.

There are several deficiencies of the existing keyword search techniques for retrieving verses of the Quran (Aya'at). These are: some irrelevant verses are retrieved; some relevant verses are not retrieved; or, the sequence of retrieved verses is not in the correct order (Shoaib et al., 2009). The keyword-based techniques have additional limitations, including misunderstanding the exact meaning of the input words forming a query, and neglecting some IR techniques(Raza, Rehan, Farooq, Ahsan & Khan, 2014).

Moreover, the current Quran semantic search techniques also have limitations concerning their ability to find the requested information. These constraints result in ambiguity in the results because semantic search tools use one or two incomplete Quranic ontologies and ignore the others. Additionally, these ontologies have different scopes and formats that require alignment and normalisation (Alrehaili & Atwell, 2014).

This thesis will evaluate the existing search tools developed for the Holy Quran against 13 criteria related to: search features, output features, the precision of the retrieved verses, recall database size, and types of database content.

Then, the study will review existing Quranic ontologies and compare them against 11 criteria. Some deficits have been identified in all these ontologies. Additionally, no single Quranic ontology covers most of the knowledge contained in the Quran. Therefore, a new Arabic-English Quran ontology was developed from ten datasets related to the Quran, such as: Quran chapter and verse names; word meanings in the Quran; and topics in the Quran. The main aim of developing a Quranic ontology is to facilitate the retrieval of knowledge from the Quran. Additionally, the proposed Quranic ontology will enrich the raw Arabic and English text of the Quran with Islamic semantic tags.

Furthermore, for this study the first Annotated Corpus of Quran Questions and Answers in Arabic was developed. This corpus contains 2,200 question-answer pairs collected from trusted Islamic sources. Each question-answer pair is labelled with five tags. these tags are: a question type: either factoid or descriptive; a topic of question based on Quranic ontology; and, a class of the question.

Finally, the thesis proposes a new semantic search model for the Arabic Quran based on the developed Quranic ontology. This model aims at overcoming the limitations of the existing Quran search applications. The search tool employs both x techniques and semantic search technologies. The performance of this search model is then evaluated using the Annotated Corpus of Arabic Quran Questions and Answers.

## 1.2 Motivation

The primary motivation for this study is to enrich the raw Arabic Quran text with Islamic ontology that could help the reader to develop a better understanding of the Quran. Additionally, the Quranic ontology can be used to understand the meaning of any query relating to the Quran and classify retrieved results based on this ontology.

The study will focus on how to address the ambiguity in search queries and retrieved results using a hybrid of both semantic search and keyword-based techniques based on Quranic ontology.

The motivation for selecting the Quran as the subject of semantic search can be summarised in the following points:

- The Quran is the leading resource for classical Arabic language.
- The Quran discusses various topics, for example, ethics, Islamic law, marital and family law, monetary transactions, morals, and the relationship between Islam/Muslims and other world religions.
- The above features provide a rich environment for semantic search.
- Existing Quranic ontologies in both Arabic and English languages will help to construct an Arabic semantic search based on ontology alignment, where there are limited Arabic ontology resources available for research purposes.
- The Quran is the most widely read Arabic book, not only in the Arabic regions, but also in the Islamic world. Additionally, the Quran is taught as a compulsory subject in Islamic education systems.

- Intensive computational research has been conducted in this domain, such as Arabic Quran corpus research (Dukes, 2013), mining the Quran (Sharaf & Atwell, 2012), and ontological learning from Quranic text (Alrehaili & Atwell, 2014).

## 1.3 Research problems

The following research problems were identified based on a review of the literature on the existing Quran search applications, and ongoing research:

- The limitations of existing Quran search tools for retrieving more than one verse for one query as one answer. For example, if the query is ‘What are names for Heaven in the Quran?’ “ماهي اسماء الجنة؟”, the answer would appear as in table 1.

**Table 1: Names of Heaven in the Quran**

| |
|---|
| • الحسنى، (the best reward)<br>قال تعالى: (لِلَّذِينَ أَحْسَنُوا الْحُسْنَىٰ وَزِيَادَةٌ ۖ وَلَا يَرْهَقُ وُجُوهَهُمْ قَتَرٌ وَلَا ذِلَّةٌ ۚ أُولَٰئِكَ أَصْحَابُ الْجَنَّةِ ۖ هُمْ فِيهَا خَالِدُونَ) [يونس: 26].<br>(For them who have done good is the best [reward] and extra. No darkness will cover their faces, nor humiliation. Those are companions of Paradise; they will abide therein eternally) [chapter: Yunus; verse:26]<br>• دار السلام (the Home of Peace)<br>قال تعالى: (لَهُمْ دَارُ السَّلَامِ عِنْدَ رَبِّهِمْ وَهُوَ وَلِيُّهُمْ بِمَا كَانُوا يَعْمَلُونَ) [الأنعام:127]<br>(For them will be the Home of Peace with their Lord. And He will be .their protecting friend because of what they used to do)<br>• جنات عدن: (Garden of Eden)<br>قال تعالى: (وَعَدَ اللَّهُ الْمُؤْمِنِينَ وَالْمُؤْمِنَاتِ جَنَّاتٍ تَجْرِي مِنْ تَحْتِهَا الْأَنْهَارُ خَالِدِينَ فِيهَا وَمَسَاكِنَ طَيِّبَةً فِي جَنَّاتِ عَدْنٍ) [التوبة: 72].<br>(Allah has promised the believing men and believing women gardens beneath which rivers flow, wherein they abide eternally, and pleasant dwellings in |

gardens of perpetual residence; but approval from Allah is greater. It is that which is the great attainment.)

- جنات النعيم، قال تعالى: (إِنَّ الَّذِينَ آمَنُوا وَعَمِلُوا الصَّـــالِحَاتِ يَهْدِيهِمْ رَبُّهُمْ بِإِيمَانِهِمْ تَجْرِي مِنْ تَحْتِهِمُ الْأَنْهَارُ فِي جَنَّاتِ النَّعِيمِ) [يونس:9].
- دار المتقين، قال تعالى: (وَلَنِعْمَ دَارُ الْمُتَّقِينَ) [النحل:30].
- جنات الفردوس، قال تعالى: (إِنَّ الَّذِينَ آمَنُوا وَعَمِلُوا الصَّـالِحَاتِ كَانَتْ لَهُمْ جَنَّاتُ الْفِرْدَوْسِ نُزُلاً) [الكهف:107].
- جنة الخلد، قال تعالى: (قُلْ أَذَلِكَ خَيْرٌ أَمْ جَنَّةُ الْخُلْدِ الَّتِي وُعِدَ الْمُتَّقُونَ كَانَتْ لَهُمْ جَزَاءً وَمَصِيراً) [الفرقان:15].
- الغرفة، قال تعالى: (أُولَئِكَ يُجْزَوْنَ الْغُرْفَةَ بِمَا صَـبَرُوا وَيُلَقَّوْنَ فِيهَا تَحِيَّةً وَسَـلاماً) [الفرقان:75].
- دار المُقامة، قال تعالى: (الَّذِي أَحَلَّنَا دَارَ الْمُقَامَةِ مِنْ فَضْلِهِ لا يَمَسُّنَا فِيهَا نَصَبٌ وَلا يَمَسُّنَا فِيهَا لُغُوبٌ) [فاطر:35].
- دار القرار، قال تعالى: (يَا قَوْمِ إِنَّمَا هَذِهِ الْحَيَاةُ الدُّنْيَا مَتَاعٌ وَإِنَّ الْآخِرَةَ هِيَ دَارُ الْقَرَارِ) [غافر:39].

- Most search tools do not analyse and classify the query texts by applying natural language processing (NLP) and semantic techniques. Examples of NLP analyses include Part of Speech (POS) tagging, named entities recognition (NER), parsing, and spelling corrections. An example of a semantic technique is using an ontology to disambiguate the conceptual meaning of a user query.

- The existing Quranic ontologies have different scopes and formats. Therefore, most Quranic search tools only use one source of Quranic ontology. Additionally, some Quranic ontologies are not available for use. In other words, Quranic ontologies lack the advantages of ontology development. These advantages include enabling the sharing and re-use of knowledge. Additionally, the better engineering of Quran ontology will enhance and ease the process of maintaining and expanding Quranic ontology.

- The NER of the Arabic language is mostly focused on the modern Arabic language. Additionally, the sets of Arabic NE do not cover the classical

Arabic language words. Furthermore, no well-formatted NE lists exist that are specialised for Quranic text, such as those of Allah's names, the Prophet's names, lists of animals, times, religions, and so on.

- Arabic is a highly inflected language with a complex orthography. This will increase ambiguity in the search results.

Unlike English language research datasets, there are lack of Arabic and Islamic annotated question-answer datasets for testing and evaluation of Islamic-Arabic search techniques. Therefore, most research conducted in IR in the Quran domain has produced imprecise evaluations.

## 1.4Research aims and objectives

The aim of the study is to develop a new Arabic Quranic semantic search model based on a new Quranic ontology. This model will use advanced hybrid semantic and information retrieval techniques. In addition, the model returns verses from the Quran as an answer to a user's query. Moreover, the model will consist of new resources: the new Quranic ontology, the annotated Arabic Quranic question-answer corpus, and the Arabic Quranic question classifiers.

Researchers interested in computational linguistics might use this model and the newly developed resources as valuable resources to apply to another research areas of Arabic language including natural language processing, information retrieval, knowledge extraction and Islamic study. Furthermore, other potential users of this tool could be general users and Islamic scholar who might benefit from this project in teaching, learning and exploring knowledge in the Quran.

To overcome the research problems discussed in the previous section, this study pursues several objectives:

1 Evaluate the existing search tools that have been constructed for the Holy Quran.

2 Review existing Quranic ontologies and compare them to ontology development methodologies. Then, develop a new Arabic-English Quranic ontology using existing datasets related to the Quranic sciences.

3 Develop an Annotated Corpus of Quran Questions and Answers in Arabic. This corpus will contain question and answer pairs collected from trusted Islamic sources. Each question and answer pair will be labelled with question-type tags.

4 Develop a new semantic search model for the Arabic Quran based on the developed Quranic ontology. This search model will employ both IR techniques and semantic search technologies. The performance of this search model will be evaluated using the developed Annotated Corpus of Arabic Quran Questions and Answers.

## 1.5 Research questions

This research aims to provide answers to the following questions:

**Will the new Arabic Quranic semantic search model based on the Quranic ontology reduce the ambiguity in the retrieved results?**

The core aim of semantic search is to understand the user's query and return a disambiguated result (Amerland, 2013). The main challenges in the previous research on retrieving knowledge from the Quran have been resolving the ambiguities in query understanding, which affects the answers given in response to the query. Figure 1 demonstrates the word sense disambiguation of the word 'eljannah'.

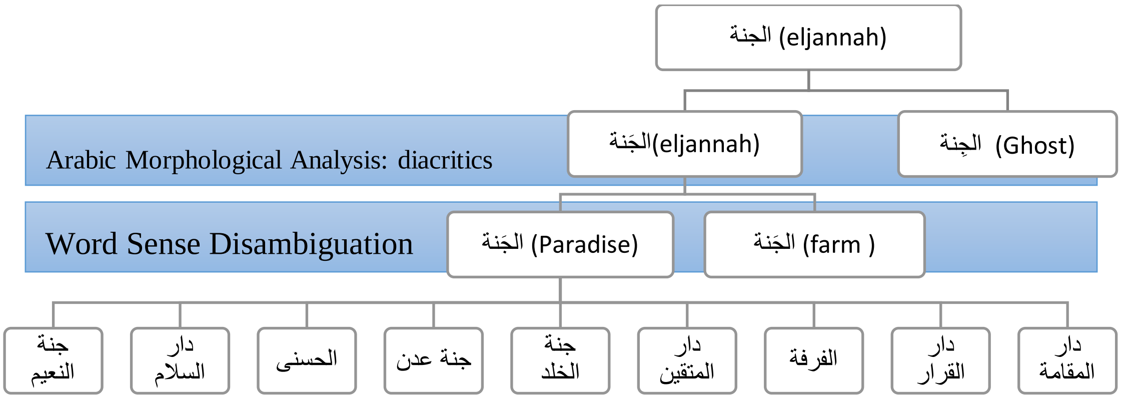


Figure 1: Arabic morphological analysis and Word Sense Disambiguation of word 'eljannah' الجنة in the Quran Domain

In previous research on search tools for the Quran, only one Quran dataset has been used as a source of ontology for these search tools (Mohammad Alqahtani & Atwell, 2017). This could affect the accuracy of the search results because ontology does not cover all the concepts mentioned in the Quran. Consequently, the Quranic ontologies could be built from many Islamic datasets, to cover most of the varied types of knowledge in the domain of the Quran. A semantic search tool based on this new Quranic ontology could be reflected in the accuracy of retrieved results of a user's query in terms of the most relevant answers to a query.

**Will developing a Quranic classifier using Quranic ontology classes, enhance the prediction of answer types?**

Question classification is used to deduct an answer type for a user's query to reduce the size of searchable documents and increase the accuracy of the retrieved results. Therefore, classifying queries as entities, might increase the accuracy of the retrieved results. However, Quranic question classification is relaying on limited classes including person, animal, entity, description, location and other (X. Li & Dan, 2002).

This question could be answered by constructing a question hierarchy relevant to Islamic domain. Then, developing a new model for Arabic Quranic classifier using deep learning methods of text classifications. After that, this classifier needs to be trained and examined on an annotated Quranic question dataset. Finally, this testing results could be evaluated using appropriate IR evaluation metrics.

## 1.6 Research contributions

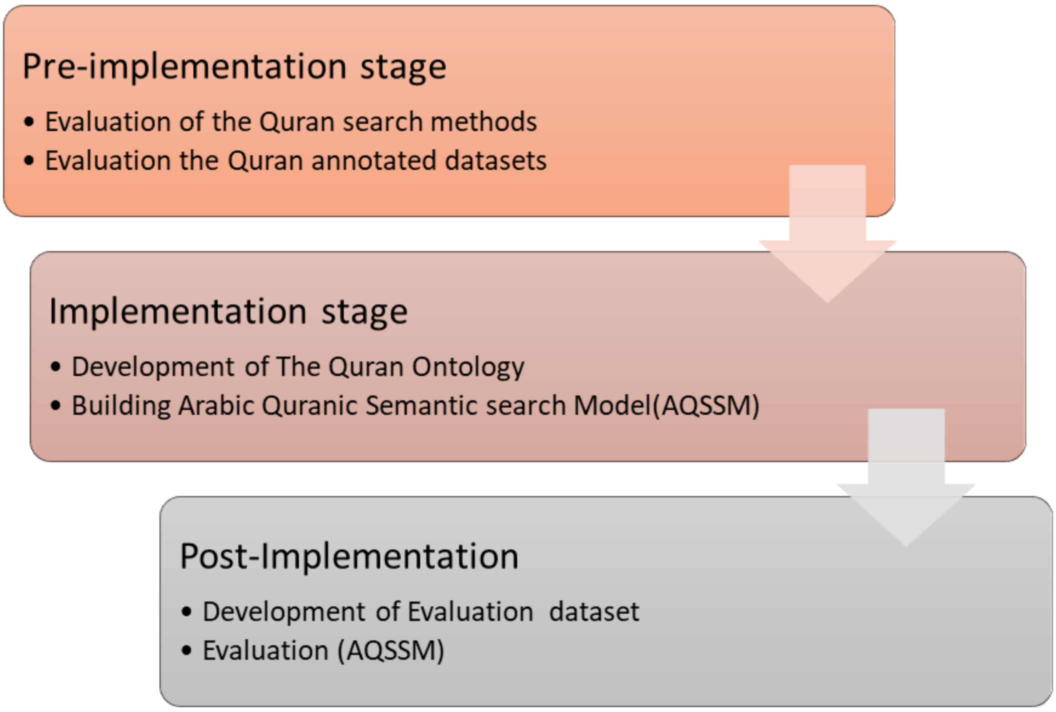
The contributions and originality of this research emerge from the points below; these are based on each of the objectives listed above, the achievement of which represents a research contribution:

1. Evaluation criteria for existing search tools constructed for the Holy Quran.
2. Review the existing Quranic ontologies and the subsequent development of a new Arabic-English Quran ontology that combines existing datasets related to the Quran.

3. The development of the first Annotated Corpus of Quran Questions and Answers in Arabic collected from trusted Islamic sources.
4. The development of a new semantic search model for the Arabic Quran based on the developed Quranic ontology.
5. An additional contribution of this research is a new Arabic question classifier for Islamic questions.

## 1.7 Research Methodology

This section outlines the methodology of implementing the proposed Arabic Quranic semantic search Model (AQSSM). The methodology followed to implement this study three stages: pre-implementation stage, implementation stage and post-implementation stage, as shown in figure 2.



Pre-implementation stage
- Evaluation of the Quran search methods
- Evaluation the Quran annotated datasets

Implementation stage
- Development of The Quran Ontology
- Building Arabic Quranic Semantic search Model(AQSSM)

Post-Implementation
- Development of Evaluation  dataset
- Evaluation (AQSSM)

**Figure 2: Overview of the research methodology**

Firstly, the pre-implementation stage evaluates the existing methods and resources for the Quran. In this phase, two sets of evaluation criteria were designed using the best common evaluation practices in the information retrieval systems and knowledge representation. The first group of evaluation criteria has 14 different measures to assess the present Quranic search methods applied to the Quran (Chapter 3 will explain this in detail). The second group of evaluation criteria consists of 13 different measures to judge the previous Quranic annotated datasets. The ontology evaluation criteria were designed using best practises for evaluating ontology (Chapter 4 will discuss this in detail).

Secondly, the implementation stage consists of 2 components: developing the Quran ontology and constructing AQSSM. The development of the new Quranic ontology followed common standards from the different methodologies used for ontology development, as discussed in chapter 4. The development of AQSSM consists of deep learning and statistical machine learning methods, including word2vec and fastText, Vector space model.

Finally, the post-implementation stage aimed at evaluating the AQSSM in order to examine the performance of this model. Therefore, the AQSSM was evaluated by using the most appropriate IR evaluation measures and a specific testing dataset. Chapter 5 will describe the testing dataset, and chapter 6 will explain the evaluation metrics used to assess the testing results of AQSSM.

## 1.8 Thesis organisation

The remainder of this thesis is divided into seven chapters:

Chapter One has summarised the research problems addressed and explained the field of study. Moreover, it has elaborated on the motivations and rationale for choosing the Quran as the topic for this PhD project. Then, the scope of the project was explained, and the novelty and original contributions were highlighted. Finally, the structure of the thesis was outlined.

Chapter Two will provide a general background to information retrieval (IR), the Semantic Web, and ontology development. It will then present a literature review with a focus on work related to religious search applications and methods.

Chapter Three will review the search tools constructed for IR from the Holy Quran. Additionally, this chapter will evaluate these different search tools against

14 criteria relating to search features, output features, the precision of the retrieved verses, recall, database size, and types of database content.

Chapter Four will review the existing Quranic ontologies. Additionally, the chapter will describe the eleven criteria to evaluate the Quranic ontologies. Then, the chapter will summarise the evaluation results of the Quranic ontologies. Finally, the chapter will describe the development and evaluation process of the new Arabic-English Quranic ontology.

Chapter Five will provide some background on corpus linguistics. Furthermore, it will describe the different types of corpus linguistics. Moreover, the chapter will emphasise the importance of Question-Answer corpora and summarise their use in IR systems. It will then detail the construction stages of the new Annotated Arabic Quran Question and Answer Corpus (AAQQAC). After that, the chapter will describe the new classifier for the Arabic questions related to the Quran. Finally, the chapter will introduce a semantic similarity model for Arabic words in the domain of the Quran using word embedding techniques.

Chapter Six will present a new search model called a Semantic Search Model for the Quran based on Quranic ontologies. Additionally, the chapter will explain in details the framework of this model. Finally, the chapter will discuss the evaluation result of the model using the IR evaluation measures.

Chapter Seven will summarise the contributions of this thesis. It will also outline some plans for future work on each component of this project. The chapter will discuss the challenges faced in the study and the limitations that necessitate further work, before presenting the conclusions drawn from this experimental work.

# Chapter 2
# Background

## 2.1 Introduction

Chapter two will provide a general background to information retrieval, the semantic web, and ontology development. It will then present a literature review with a focus on work related to religious text search applications and methods.

## 2.2 Information retrieval

### 2.2.1 Definition of Information retrieval

Information retrieval (IR) can be defined as obtaining information that meets an information need from large collections of text, usually stored on computers. In the 1945 article *As We May Think*, Vannevar Bush suggested using computers to search for specific information (Singhal, 2001). Not long after, in the 1950s, the first automated IR systems were announced. Most people engage in IR daily when they use World Wide Web search engines, such as Google or Yahoo. The primary goals of IR are to enable users to search vast collections of data quickly, to provide more flexible matching algorithms including stemming query words, and to enable the ranking of retrieved results.

### 2.2.2 IR techniques

IR systems use many techniques to increase the quality and precision of search results. Examples of these methods are inverted indexes, pre-processing before document indexing, sorting results, term weighting, the vector space model, and evaluation of an IR system based on the relevance of the documents it retrieves. However, ambiguity in the search result is still a significant issue in IR.

The new version of IR is called semantic search. This search works by reasoning about an input query over a knowledge base and returning the most relevant answers. The input query can have different forms such as a natural language question, a triple knowledge representation of a question, a graphical representation, or keywords. The knowledge base can be one or more ontologies, corpora or plain text documents. Similarly, the retrieved answers from semantic search can take a multitude of forms from pure triples to a natural language representation.

### 2.2.3 IR systems evaluation

Evaluation metrics in IR systems aim to evaluate how well search results match the intent of the query. The most popular evaluation measures used for IR systems are: recall, precision, F-measure, precision at k, recall at k and average precision. The recall, precision and F-measure are used to measure the effectiveness of the unordered retrieved set (Zhu, 2004).

The recall is the percentage of retrieved relevant documents:

$$recall = \frac{\#(retrieved\ relevant\ documents)}{\#(relevant\ documents)}$$

Precision is the proportion of retrieved items that are relevant:

$$Precision\ = \frac{\#(retrieved\ relevant\ documents)}{\#(retrieved\ documents)}$$

On the one hand, recall assesses the system's ability to find all relevant documents. On the other hand, precision evaluates the system's ability to discard any non-related documents in the retrieved documents. Therefore, F-measure is used to measure the balance between precision and recall by determining the harmonic mean of both precision and recall:

$$F - measure\ = \frac{2\ \times\ precision\ \times\ recall}{precision +\ recall}$$

Precision at k, recall at k and average precision could be used to assess the order of the retrieved documents (Manning et al., 2008).

Precision at k is the percentage of recommended items in the top-k set that are relevant.

$$precision\ @\ k = \frac{\#(recommended\ items\ @k\ that\ are\ relevant)}{\#(recommended\ items\ @k\ )}$$

where recall at k is the fraction of relevant items in the top-k of the retrieved set.

$$recall\ @\ k = \frac{\#(recommended\ items\ @k\ that\ are\ relevant)}{\#(\ relevant\ items\ )}$$

Using precision and recall at *k* to rank document retrieval make the process of computing and interpreting the metrics simpler. However, the drawbacks are that the value of k has a significant effect on the metric, and any ranking above *k* is inconsequential. In other words, no rule exists to determine the best value of k. Therefore, to overcome the issue of choosing the k-value to calculate precision, the average precision metric can be used to calculate precision and recall without having to set k value.

$$average\ precision = \frac{\sum_{k=1}^{n}(P(k) \times\ rel(k))}{\#(\ relevant\ items\ )}$$

Where:

*n* indicates the number of retrieved documents;
*k* is the position in the sequence of retrieved documents;
P(k) is the value of precision at k in the list;
and *rel(k)* equals 1 if the item at rank k is a relevant document, zero if not.

This average precision can be used for a single query. The average precision across a set of queries is called the mean average precision (MAP).

$$MAP = \frac{\sum_{q=1}^{\#Queries} Average\_Precision(q)}{\#\ Queries}$$

In this research, the proposed Arabic Quranic semantic search model returns a ranked retrieval result for a query. Therefore, average recall, average precision and MPA could be the best measures for evaluating ranked retrieval results (further explanation in section 6.3.1).

## 2.2.4 Question & Answers classification

This section will discuss the two paradigms of Question and Answering techniques. Furthermore, the section will discuss the methodologies of classifying questions.

### 2.2.4.1 Question and Answering paradigms

In the1960s, ‘knowledge-based question answering’ and ‘Information Retrieval (IR) based question answering’ were the main paradigms of Question and

Answering schemes implemented by searching systems. IR-based question-answering aims to answer a question by identifying a relevant short text segment in a collection of documents. The knowledge-based question-answering approach to answering a human question is by mapping this question to a query over a structured database. An example of the knowledge-based question- answering approach is the BASEBALL system which belongs to the text-based paradigm for question answering. This system was created in the beginning of the 1960s which is the start of the revolution of natural language processing (Green et al., 1961).

### 2.2.4.2 Question classification

Questions are classified based on their answers into two types: detailed and factual questions. The factual questions are defined as questions that can be answered with simple facts expressed in short text answers, while the detailed questions are answered with description or explanation.

The process of question classification is defined as the task to determine the answer category or the named-entity (Jurafsky & Martin, 2017). For instance, a question such as “Where is the Holy Mosque?” expects an answer category ‘LOCATION’. Therefore, the answer category or “answer type recognition” depends on a set of named-entity hierarchy that is referred to as the answer type taxonomy. This taxonomy can be automatically constructed from Word-net (Paşca, 2003) or manually built (X. Li & Dan, 2002). For instance, the first Named Entity set has seven classes: location, date, organisation, money, per cent, person, expressions, and time. Li and Dan (2002) developed two additional layers of question hierarchy: the first layer has 6 coarse classes and the second has 50 fine classes.

There are two approaches to classifying questions: the rule-based approach and the machine learning approach. The rule-based approach is a set of written rules that are used to predict the answer type. For example, the expected answer type of a ‘When’ question is time. On the other hand, the anticipated answer types might be derived through the machine learning approach. This technique relies on three parts: the taxonomy of answer types into which questions are to be classified; a corpus of annotated questions with the associated answer type classification; and, an algorithm that is trained to make the actual predictions given this corpus (Jurafsky & Martin, 2017).

In this study, the machine learning approach is used to classify questions about the Quran. This will be described in section 5.5.

## 2.3 Semantic web

The Semantic Web (SW) is a web of data providing a common framework that permits data to be shared and reused across applications. Tim Berners-Lee, the initiator of SW, defines it as “an extension of the existing web in which information is awarded well-defined meaning, well enabling computers and people to work in co-operation” (Shadbolt, Hall, & Berners-Lee, 2006). SW is based on the Resource Description Framework (RDF) for representing knowledge (Hitzler, Krotzsch, & Rudolph, 2009) . The main goals of SW are to make data understandable by computers and humans, and to represent knowledge as linked data. In an SW context, the relationships between concepts are called vocabularies, or terms. These terms are used to define and characterise an area of concern. From a practical perspective, vocabularies are more complicated when they contain thousands of terms, or they can be very simple when they describe just a few concepts. Complex vocabularies are called ontology(W3C, 2015). SW uses different languages to demonstrate data as graphs, (i.e. RDF), explaining this data organisation via ontology, (i.e. OWL), and querying it (i.e. SPARQL).

## 2.4 Semantic search

A semantic search is an application of SW that has shown significant potential for improving the performance of retrieval. Compared to traditional search engines that emphasise the occurrence of words, the semantic search engine attempts to understand the meanings of queried terms and, based on this, retrieve documents with matching meanings. This might be achieved by adding semantic tags to texts in order to structuralise and conceptualise the objects within documents (Dong, Hussain, & Chang, 2008). In other words, a semantic search uses SW technologies to perform search and IR activities.

Traditional search engines, such as Google, Yahoo, and Bing currently apply many semantic search approaches. These techniques allow the search engines to dominate the search engine market. As an illustration, Google began to use semantic search techniques by introducing a Google Knowledge Graph[1]. This graph is designed to enhance search results based on the linked information gathered from a wide variety of sources, such as the *CIA World Factbook*[2], Freebase[3], and Wikipedia[4]. The primary purpose of the knowledge graph is to develop a question-and-answer engine, such as Ask Jeeves[5] or Wolfram Alpha[6]. Currently, many semantic search engines enable users to explore the web by topic, relevant media, opinion, communities, and links to related topics. Hakia [7] is an example of a semantic search engine that returns relevant results based on concept-matching rather than keyword-matching; this engine allows a user to enter keywords, a phrase, or a question. Semantic search engines can be divided according to the four approaches used: contextual analysis, reasoning, natural language analysis, and knowledge representation using ontology(Manning et al., 2008).

### 2.4.1 Semantic search types

Existing research approaches in the field of semantic search can be categorised into two types: approaches based on Structured Query Language (SQL), and approaches for naïve users (Fazzingaa & Lukasiewiczb, 2010).

In the semantic search systems using the SQL approach, the user can navigate to a particular topic through an ontology graph or use an SQL such as SPARQL to retrieve a particular subject from an RDF document. For example, Swoogle provides a search function for Semantic Web documents and terms primarily by indexing and querying RDF documents; the results are the URIs of classes. A URI (Uniform Resource Identifier) is a string which refers to a resource in the web.

[1] *https://www.google.com/insidesearch/features/search/knowledge.html*
[2] *http://en.wikipedia.org/wiki/CIA_World_Factbook*
[3] *http://en.wikipedia.org/wiki/Freebase*
[4] *http://en.wikipedia.org/wiki/Wikipedia*
[5] *http://en.wikipedia.org/wiki/Ask_Jeeves*
[6] *http://en.wikipedia.org/wiki/Wolfram_Alpha*
[7] *http://www.hakia.com/*

The most well-known is URLs, which identify the resource by giving its location on the Web (Yu, 2014).

By contrast, systems using the approach, for naïve users, prompt users to specify their queries without demanding any knowledge about ontologies or specific query languages. The approaches used in such systems can be divided into two broad methodologies. First, keyword-based approaches, where a query contains a list of keywords. For instance, the user's query could be "disease", "chronic", and "the UK". Another methodology consists of natural-language-based approaches, where a query takes the form of a natural language sentence such as, "The most common risk factors causing chronic diseases in Europe". SemSearch is an example of a natural-language-based search system; query keywords are given semantic meaning by matching them with similar classes, properties, or instances from the semantic database. These semantic meanings form a query that is used for retrieving results that are semantically related to all the user keywords (Lei, Uren, & Motta, 2006).

## 2.5 Ontology

### 2.5.1 Definition

In computer science, the term 'ontology' is defined as an explicit specification of a conceptualisation of a domain, in terms of concepts, attributes, and relations (Gruber, 2009). Common components of ontologies are: classes (concepts), attributes, relations, functional terms, restrictions and axiom. Concepts here are the entities of interest in a specific domain. These concepts are structured into a taxonomy tree or un-taxonomy tree. Each tree node represents a concept that is a specialisation of its ancestor. The concept is also related to a set of instances and has a set of attributes. Relations are the ways in which concepts and instances can be linked to each other.

### 2.5.2 Developing ontology

The main goals of developing an ontology are to share a common understanding of the structure of concepts between people and software agents, and permit reuse of domain knowledge (Gruber, 1993).

In computer science, there are two types of ontology: a domain ontology and upper ontology.  Domain ontology is also referred to as domain-specific, which represents a group of concepts describing a particular domain. For example, the meaning of the word 'card' in the domain of poker differs from its meaning in the domain of computer hardware. Domain ontology is employed in restricted-domain question-answer systems to formalise domain knowledge and represent natural language questions and underlying unstructured information sources.
By contrast, upper ontology explains common concepts that are mostly applicable across several domains. This type of ontology is available for public use, for example WordNet. Typically, some of the current upper ontologies are exploited to complement the domain-specific ontologies to enhance the available domain resources among semantic connections and definitions.

In a Semantic Web context, many languages have been proposed to develop ontologies, such as RDF[8], OIL[9], DAML[10], DAML+OIL[11], and OWL (Gruber, 2009; Ou, Pekar, Orasan, Spurk, & Negri, 2008; W3C, 2015).

There is no standard methodology for ontology development. However, several outstanding methodologies and best practices exist for constructing an ontology. Some examples of methodologies include: On-To-Knowledge (Staab & Studer, 2001), OntoSpec[12] (Kassel, 2005), the NeON[13] methodology (Gómez-Pérez & Suárez-Figueroa, 2008), MOKI (Ghidini et al., 2009), the Melting Point methodology (Garcia et al., 2010), and DiDOn (Keet, 2012).

*8 https://www.w3.org/RDF/*

*9 Ontology Interchange Language (OIL): is depend on concepts developed in frame-based systems, and Description Logic. OIL is compatible with RDFS.*

*10 DAML: stands for The DARPA Agent Markup Language Homepage. More details on http://www.daml.org/*

*11 DAML+OIL is a semantic markup language for Web resources. it can provide a rich set of constructs to create ontologies and to markup information which is readable by computers*

*12 OntoSpec is a service that automatically extracts classes, object properties, data properties, and namespace declarations from an OWL and OWL2 ontology, and renders them as ordered lists, together with their textual definitions, in a human-readable HTML page designed for browsing and navigation by means of embedded links. http://ontospec.com/*

*13 A network of ontology (NeOn) is a collection of related ontologies via a different relationships such as  modularization, mapping, dependency relationships,  and version*

The life cycle stages shared by most methodologies for ontology development are: specification, conceptualisation, formalisation, implementation, evaluation, and documentation.

The specification stage determines the ontology's purpose and scope. The conceptualisation stage identifies the concepts to be included in the ontology, and how they relate to each other; this will be dependent, to some extent, on the ontology's scope and competency questions. Therefore, classes and relationships must have exact names and descriptions.

The formalisation stage is when the hierarchy of concepts and relations is determined. Additionally, it is where any constraints are defined, such as 'concept A is_Disjoint_from concept B', has unique email, and has a maximum number. At this stage, the most popular approaches are used to construct the class hierarchy are: top-down, bottom-up, or middle-out. On the one hand, the top-down approach begins with the most general classes and ends with the most detailed classes. On the other hand, the bottom-up approach starts with the most specific classes and ends with the most general classes. The middle-out approach begins with the most important classes; then, this method moves upward to the more general, and downward to the more specific classes. The middle-out approach is an excellent method for controlling the scope and details of classes. In addition to the formalisation stage, naming conventions for ontology components are not compulsory but strongly recommended. The name convention is a set of rules for writing the names of elements of the ontology. For example, spaces and uncommon delimiters must be avoided in class and relation names. Additionally, the first letter in each word in the class name should be capitalised; however, relation names generally starts with a lowercase letter.

After the formalisation stage, the ontology could be implemented with one of the ontology development tools such as Protégé. At this stage, an appropriate ontology language might be chosen based on the type of ontology; RDFS, OWL Lite, OWL, and DL are examples of ontology languages. During the implementation stage, an ontology tool performs many processes, such as: editing the class hierarchy, adding relationships and restrictions, selecting appropriate value types, cardinality, and applying a reasoner. The reasoner checks the consistency and satisfiability of this ontology. Examples of a reasoner include Fact++[14], Racer[15], and Pellet[16].

The evaluation stage tests the validity of the ontology in terms of syntax, competency, and efficiency. The ontology can be further validated using online tools such as W3C RDF validator[17]. Its efficiency can be measured using two factors, time and accuracy when answering a query.

The final stage of the ontology development lifecycle is documentation. This stage involves documenting, in detail, the design options, assumptions, structure decisions, and examples. Skuce (1995) suggested a format for documenting ontological conventions; this format includes class and relation assumptions, conceptual assumptions, terminological assumptions, definitional assumption, and examples. This documentation is highly important for future usability and understanding of the ontology.

The developed ontology can be stored in a Graph Database. The Graph Database is a subject-predicate-object database server (triple store). This is used to provide the protocol engine for other RDF query and storage systems. Apache Jena Fuseki[18], GraphDB[19] and Neo4j[20] are examples of the graph database systems. For example, the Neo4j has more features than Fuseki such as graphical presentation of concepts, powerful query language called Cypher, and API with many programming languages such as python and PHP. A DB-Engines Ranking [21] ranks graph database management systems (DBMS) rendering their popularity. This rank is updated every month.

Current challenges faced in ontology development are over-scaling and complicating the ontology, lack of documentation, redundancy, and using

[14] *FaCT++ is the new generation of the well-known FaCT OWL-DL reasoner. More details at http://owl.man.ac.uk/factplusplus/*

[15] Racer is a knowledge representation system that implements a highly optimized tableau calculus for the description logic . http://www.ifis.uni-luebeck.de/~moeller/racer/

[16] Pellet is an open source OWL Deep Learning reasoner work with Java.

[17] *https://www.w3.org/RDF/Validator/*

[18] *https://jena.apache.org/documentation/serving_data/*

[19] *https://ontotext.com/products/graphdb/*

[20] *https://neo4j.com/*

[21] https://db-engines.com/en/ranking/graph+dbms

ambiguous terminology. These pitfalls increase the possibility of inconsistencies, maintenance costs, and difficulty of mapping the ontology to another ontology.

### 2.5.3 Ontology alignment

Ontology alignment (also called ontology mapping or ontology integration) can be defined as the process of finding a one-to-one correspondence between entities of two ontologies. The main goal of ontology mapping is to integrate different ontologies within the same domain (Zaeri & Nematbakhsh, 2015). The main reason for ontology mapping is that single ontology not often fulfils the needs of a specific application (Antoniou & Frank, 2012).

Recently, a large number of ontology-matching systems have been developed to identify such correspondences. Euzenat and Shvaiko (2013) classified alignment techniques into four core modules: terminological, structural, extensional, and semantic techniques.

Terminological techniques match entities based on the similarity between the names of entities. The majority of the alignment tools use terminological techniques as an initial stage. Terminological techniques are divided into string-based and language-based approaches. The string-based approach matches entities based on the similarity between letters of two words; so, for example, ‘author’ and ‘authority’ are deemed more similar than ‘author’ and ‘writer’. By contrast, the language-based technique aligns two entities that share the same meaning, for instance, ‘paper’ and ‘article’.

On the other hand, the structural approaches identify correspondences between entities depending on the internal structure of the entity and how it is connected to other entities. In other words, the structural method matches entities based on the ontology graph. Most of the existing alignment tools use terminological techniques as the initial step and then apply structural techniques to improve the outcome.

The extensional technique is based on the idea that two instances of entities will share more commonalities than differences; this leads to a higher probability of matching them.

Nevertheless, semantic techniques mostly employ theoretical models and deduction algorithms to find similarities between the clarifications of entities.

Consequently, semantic techniques are typically utilised for validation of identified correspondences.

In this study the terminological techniqe is used to align the Quran ontologies that have the same knowlage domain. This will increase the data quailty and reduce the redendancy in the newly developed ontology classes. In this technique, the Keyword match algorithim and Simple Fuzzy String Similarity algorithm are applied to find similar concepts. More details about the Quran ontologies alingment will be described in section 4.3.2.

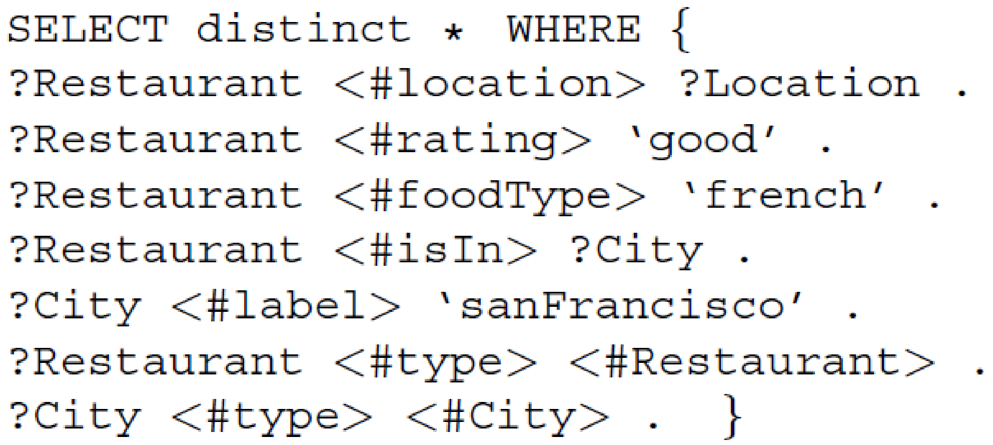
## 2.6 Semantic search based on ontology alignment

This section will provide an overview of the most recently developed semantic search tools based on ontology.

**NLP-Reduce**[22] (Kaufmann, Bernstein, & Fischer, 2007) is a natural language (NL) interface for semantic search, and uses Ginseng tags to match the RDF data (Abraham, Esther, & Christian, 2005). **NLP-Reduce** comprises five main components: a user interface (UI), an input query processor, a lexicon, a SPARQL query producer, and an ontology access layer. The UI enables the user to write an NL query as an input, and then returns a generated SPARQL query as results.

In regard to the lexicon component, all lexicons are automatically built and stored after extraction of all triples (subject-property-object) from the loaded ontology. Then, synonyms for all labels of the triples are obtained from WordNet to be used when querying. Next, the input query processor removes stop words and then matches the remaining query words to the triples stored in the lexicon. Finally, the input query processor generates a SPARQL query to retrieve the requested result.

[22] *https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/talking-to-the-semantic-web/nlpreduce/index.html*



```
SELECT distinct * WHERE {
?Restaurant <#location> ?Location .
?Restaurant <#rating> 'good' .
?Restaurant <#foodType> 'french' .
?Restaurant <#isIn> ?City .
?City <#label> 'sanFrancisco' .
?Restaurant <#type> <#Restaurant> .
?City <#type> <#City> .  }
```

**Figure 3: SPARQL code for the question "Where is a good restaurant in San Francisco that serves French food?"**(Kaufmann et al., 2007)

**FREyA** is an interactive NL interface for retrieving knowledge by querying ontologies. This system extracts possible concepts in the ontology from a user question using a syntactic parse tree. The concepts of an ontology refer to instances, classes, properties, or datatype property values, such as a date. In the case of overlaps, it issues the user a clarification dialogue to resolve the ambiguity in the concepts. After all the ontology concepts have been disambiguated and no further ontology concepts need to be determined, the system continues on to find the answer by determining the answer type. Then, this system generates a SPARQL query using combined triples of  ontology concepts (Damljanovic, 2012).

Fernández, Cantador and López (2011) proposed an IR model based on ontology. Instead of the words that appear in the documents, the inverted index stores the semantic meanings associated with the documents; these meanings are referred to as annotations. A user is allowed to submit an NL query via a question-answering tool called Power-Aqua. The system translates the NL query into ontological terms. These terms form a SPARQL query to return a list of equivalent semantic entities. Then, the inverted index is used to find a list of relevant documents. After that, this system ranks the retrieved documents by calculating their semantic similarity. The semantic similarity algorithm is a customisation of the vector space model (Castells, Fernández, & Vallet, 2007).

## 2.7 Computational search techniques for different religious texts

### 2.7.1 Introduction

The high importance of various religious texts to their believers has led to the development of IR and knowledge extraction analysis in order that these religious texts can be available online. The availability of information is a crucial factor in acquiring knowledge, and sharing information is a fundamental reason for developing ontology. This section will report the results of a survey of recently developed search tools for religious texts. This review concludes that most search applications for different holy texts are built using keyword search methods and neglecting certain IR techniques, such as indexing and scaling search results.

Religious texts, also known as scripture or holy books, are the texts that are considered to be sacred by the believers, or are central to their religious traditions. Many religions and spiritual movements believe that their sacred texts have been divinely or supernaturally revealed or inspired. This section will discuss the Bible, the Vedas, and the Tipitaka in terms of natural language processing and IR.

### 2.7.2 The Bible

The primary source of the Christian religion is the Bible, which is a collection of 66 books written by approximately 40 authors over a period of 1,600 years. The Bible encompasses many different forms of text, such as poetry, narration, fiction, history, and law.

Many desktops and web applications have been developed to retrieve information from the Bible. Examples of desktop applications are Accordance, Bible Analyzer, e-Sword, and *SwordSearcher*. Accordance[23], developed by OakTree Software, has an extensive library collection, and allows users to search by keywords, maps, and timelines. Bible Analyzer[24] is a Bible study tool with an advanced search tool named the contextual search method, which searches using more than two words in a particular context. E-Sword[25] was created by Rick

[23] *http://www.accordancebible.com*

[24] *http://www.bibleanalyzer.com/*

[25] *http://www.e-sword.net/ipad/index.htm*

Meyers, and Online Bible[26] (OLB) is a Bible reference software package established in 1987 by Larry Pierce. Both applications are computer software tools developed for studying the Bible. They offer a keyword search tool that can return verses containing the same query words.

SwordSearcher [27] (SS), developed by StudyLamp Software, is a computer programme that aims to increase Bible study amongst Christians. This application has many features and components, such as library resources and a search tool. It also has several search features and options, including the ability to find verses rapidly using words or phrases, as well as search all word forms automatically. Additionally, this tool corrects spelling errors in search words, especially proper nouns. Moreover, the user can search for related verses.

The BibleGateWay[28] (BGW) can be classified as a basic keyword search tool that enables the user to search by a keyword or phrase through either a specific or multiple Bible resources. The search results are verses that contain the exact query word/s. Additionally, this tool offers the user the opportunity to search a topic by entering keywords. However, this tool cannot retrieve verses that contain other forms of the words used in the query, for example death, die, and dying.

The BibleStudyTools[29] (BST) is a free online Bible website that allows the user to search verses by selecting from a list of Bible books and choosing a translation language, such as German, French, or Italian. The results page displays matching verses, which can be shown in context as an additional option. This tool uses only the keyword search technique.

Bibleserver.com[30] (BS) is an internet project initiated by ERF[31]. This website contains Bible resources in many different languages. The user can enter a keyword in his/her language after selecting the translation of the Bible in the same

[26] *http://onlinebible.net/*

[27] *http://www.swordsearcher.com/bible-search.html*

[28] *https://www.biblegateway.com*

[29] *http://www.biblestudytools.com/*

[30] *http://bibleserver.com/*

[31] *http://www.erf.de/online*.

language. Bibleserver.com can display results for up to four parallel Bible translations.

Biblia.com[32] is a web application that contains 50 of the books of the Bible. This website enables the user to search its library by topic or keyword. This application uses a basic keyword search.

The Blue Letter Bible[33] (BLB) website enables users to search Bible resources in two different ways: searching by word(s) or verse(s) and searching by lexical word.  In the first method, the user can enter a verse reference, such as Romans 12:1, or keywords, such as 'Jesus faith love'. After entering the search terms, the user can choose a Bible translation and then click the search button. On the results page, BLB offers the user the opportunity to see verses with all words matching the query words, as well as the option to see other verses with at least one matching query word. Additionally, BLB provides the option of seeing the different forms of the keywords being searched and the verses related to them. The second method of searching the Bible used by BLB is LexiConc, which is used to find the Greek and Hebrew terms for English words. For instance, if one runs a LexiConc search for 'love', BLB will return a list of Hebrew words (Old Testament) that are sometimes translated as 'love', followed by a Greek (New Testament) list of the same.

The Christian Classics Ethereal Library[34] (CCEL) is an electronic library containing hundreds books of the Bible. CCEL offers a search tool that uses advanced keyword search techniques for enabling full-text searches, phrase searches, stemmed searches, and searches for scripture references or commentary, as well as definitions in dictionaries or encyclopaedias. These techniques also include wildcard searches, fuzzy searches, Boolean operators, and regular expressions. The wildcard search allows the user to replace one character from the search term with '?' or '*'. For example, if the user is looking for 'test', 'tests' and 'tester', s/he can use the search term 'test*'. Fuzzy searches are based on both the Levenshtein distance and edit distance algorithms, which find the most similar words to the keyword being searched. For instance, if the user searches for 'roam'

[32] *http://biblia.com/books/esv/Jn1.1*

[33] *http://www.blueletterbible.org/*

[34] *http://www.ccel.org/*

using a fuzzy search, they will find terms such as ‘foam’ and ‘roams’. The Boolean operators allow search terms to be joined via logic operators, such as ‘AND’, ‘+’, ‘OR’, ‘NOT’ and ‘- ‘. For example, when the minus operator is used, documents that contain the term after the ‘- ‘symbol will be excluded.

STEP Bible[35] is an online application that aims to help people learn the Bible in 60 different natural languages, such as English, Arabic, German, and French. This website has many study features, such as the ability to compare different versions of the Bible, a word dictionary, original meanings and contextual information, a word cloud, and the ability to search the Bible by words, topics, and lexicons using Boolean operators and regular expressions.

**Table 2: Comparison between Bible Search tools**

| Search tool<br>Features | BGW | BST | BS | BLB | CCEL | SB | Biblia | OLB | ES | SS |
|---|---|---|---|---|---|---|---|---|---|---|
| Exact word matching | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Search for topic | ✓ |  |  |  | ✓ | ✓ | ✓ | ✓ |  | ✓ |
| Search by synonyms |  |  |  |  |  |  |  |  |  |  |
| Morphological search |  |  |  | ✓ | ✓ | ✓ |  |  |  | ✓ |
| CRLS |  |  |  | ✓ |  | ✓ |  |  |  |  |
| Boolean operators |  |  |  | ✓ | ✓ | ✓ |  |  |  | ✓ |
| Regular EX |  |  |  |  | ✓ | ✓ |  |  |  | ✓ |
| Search by finding another word form |  |  |  |  | ✓ | ✓ |  |  |  | ✓ |
| Multilingual |  | ✓ | ✓ |  |  | ✓ |  | ✓ |  |  |
| Parallel display |  |  | ✓ |  |  | ✓ |  |  |  |  |

[35] *https://www.stepbible.org*

#### 2.7.2.2 Bible ontology

SemanticBible [36]is a website that shows the ontology of the Bible, called New Testament Names (NTN). It was developed in 2006 by Sean Boisen[37].

NTN is a collection of 600 concepts extracted from names in the New Testament. Each concept is classified according to its class, such as God, Jesus, individual men and women, groups of people and locations. All concepts and their properties are defined using an RDF file. Then, NTN is represented in the Ontology Web Language (OWL) building on the RDF file and linked as can be seen in Figure 4.

[36] *http://www.semanticbible.com/*

[37] *http://seanboisen.com/*

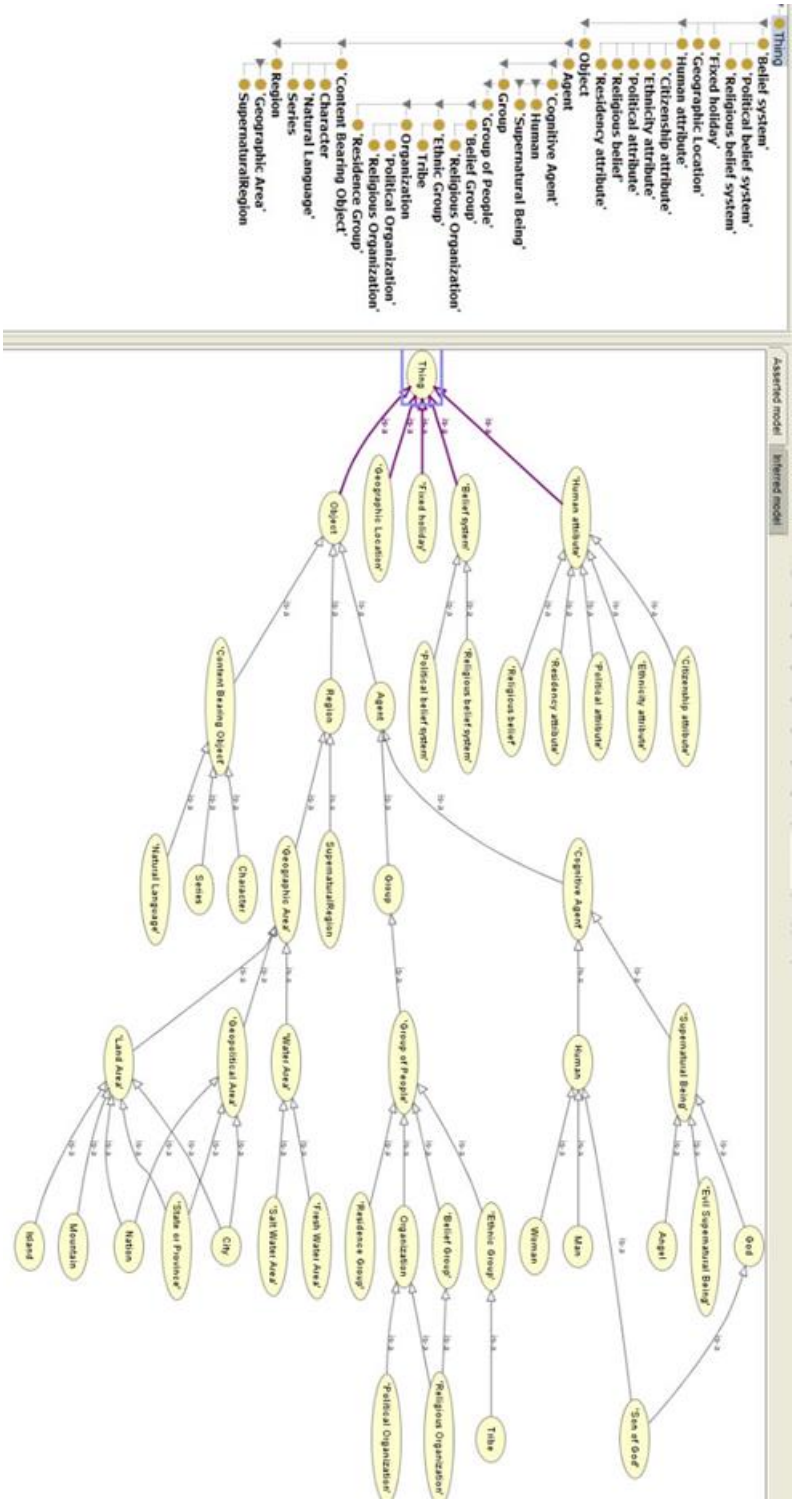
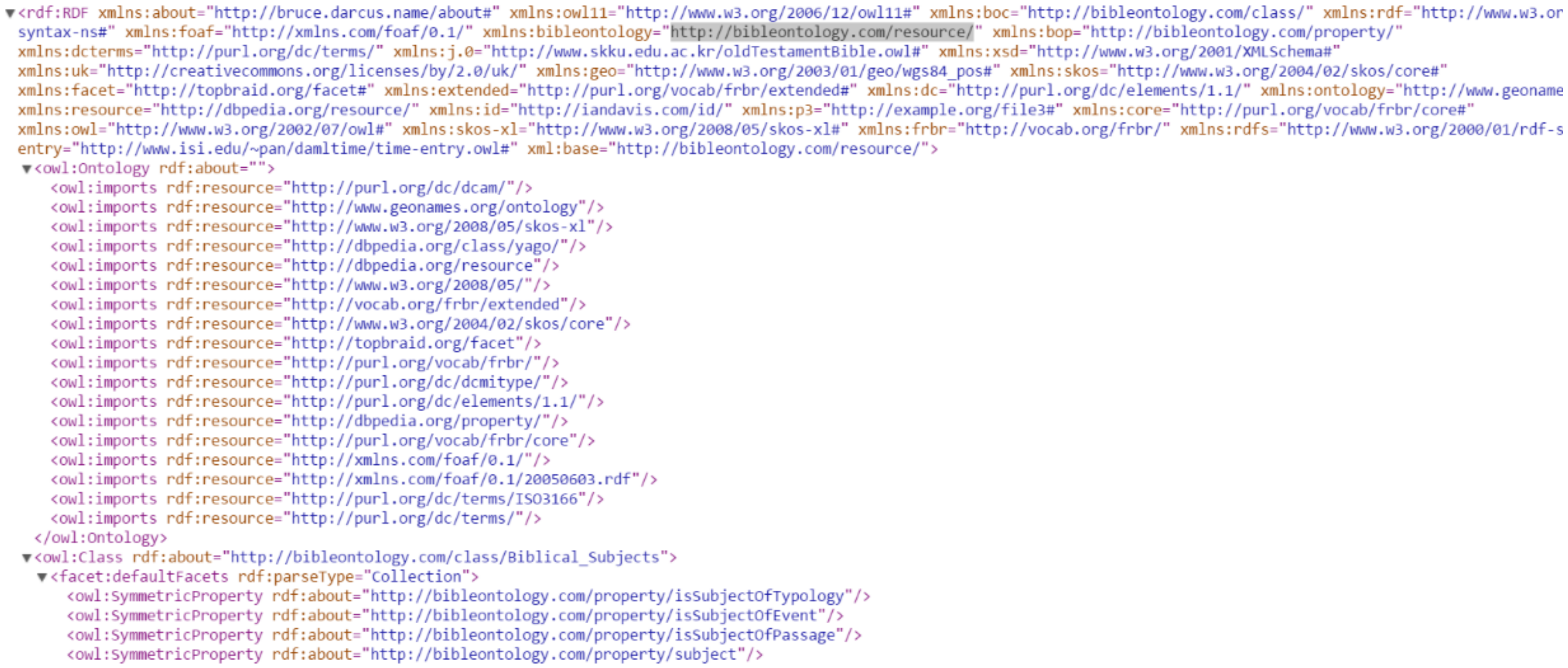


**Figure 4: Semantic Bible ontology graph**

The Bible ontology[38] was developed by Dr Myungdae Cho in 2011. This ontology contains 10,618 triples and is linked to 371 other datasets in DBPedia[39]. The Bible ontology collects all related meaningful data from multiple Bible resources and demonstrates them in the triple format, which is 'subject-predicate-object'. This dataset was built based on the SW standard. Figure 5 shows the owl file of the Bible ontology, and figure 6 demonstrates this ontology's classes.

[38] *http://bibleontology.com/*

[39] *http://dbpedia.org/*



```
▼<rdf:RDF xmlns:about="http://bruce.darcus.name/about#" xmlns:owl11="http://www.w3.org/2006/12/owl11#" xmlns:boc="http://bibleontology.com/class/" xmlns:rdf="http://www.w3.or
 syntax-ns#" xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:bibleontology="http://bibleontology.com/resource/" xmlns:bop="http://bibleontology.com/property/"
 xmlns:dcterms="http://purl.org/dc/terms/" xmlns:j.0="http://www.skku.edu.ac.kr/oldTestamentBible.owl#" xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
 xmlns:uk="http://creativecommons.org/licenses/by/2.0/uk/" xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#" xmlns:skos="http://www.w3.org/2004/02/skos/core#"
 xmlns:facet="http://topbraid.org/facet#" xmlns:extended="http://purl.org/vocab/frbr/extended#" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:ontology="http://www.geoname
 xmlns:resource="http://dbpedia.org/resource/" xmlns:id="http://iandavis.com/id/" xmlns:p3="http://example.org/file3#" xmlns:core="http://purl.org/vocab/frbr/core#"
 xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:skos-xl="http://www.w3.org/2008/05/skos-xl#" xmlns:frbr="http://vocab.org/frbr/" xmlns:rdfs="http://www.w3.org/2000/01/rdf-s
 entry="http://www.isi.edu/~pan/damltime/time-entry.owl#" xml:base="http://bibleontology.com/resource/">
 ▼<owl:Ontology rdf:about="">
    <owl:imports rdf:resource="http://purl.org/dc/dcam/"/>
    <owl:imports rdf:resource="http://www.geonames.org/ontology"/>
    <owl:imports rdf:resource="http://www.w3.org/2008/05/skos-xl"/>
    <owl:imports rdf:resource="http://dbpedia.org/class/yago/"/>
    <owl:imports rdf:resource="http://dbpedia.org/resource"/>
    <owl:imports rdf:resource="http://www.w3.org/2008/05/"/>
    <owl:imports rdf:resource="http://vocab.org/frbr/extended"/>
    <owl:imports rdf:resource="http://www.w3.org/2004/02/skos/core"/>
    <owl:imports rdf:resource="http://topbraid.org/facet"/>
    <owl:imports rdf:resource="http://purl.org/vocab/frbr/"/>
    <owl:imports rdf:resource="http://purl.org/dc/dcmitype/"/>
    <owl:imports rdf:resource="http://purl.org/dc/elements/1.1/"/>
    <owl:imports rdf:resource="http://dbpedia.org/property/"/>
    <owl:imports rdf:resource="http://purl.org/vocab/frbr/core"/>
    <owl:imports rdf:resource="http://xmlns.com/foaf/0.1/"/>
    <owl:imports rdf:resource="http://xmlns.com/foaf/0.1/20050603.rdf"/>
    <owl:imports rdf:resource="http://purl.org/dc/terms/ISO3166"/>
    <owl:imports rdf:resource="http://purl.org/dc/terms/"/>
  </owl:Ontology>
 ▼<owl:Class rdf:about="http://bibleontology.com/class/Biblical_Subjects">
   ▼<facet:defaultFacets rdf:parseType="Collection">
      <owl:SymmetricProperty rdf:about="http://bibleontology.com/property/isSubjectOfTypology"/>
      <owl:SymmetricProperty rdf:about="http://bibleontology.com/property/isSubjectOfEvent"/>
      <owl:SymmetricProperty rdf:about="http://bibleontology.com/property/isSubjectOfPassage"/>
      <owl:SymmetricProperty rdf:about="http://bibleontology.com/property/subject"/>
```

**Figure 5:The  Bible ontology**

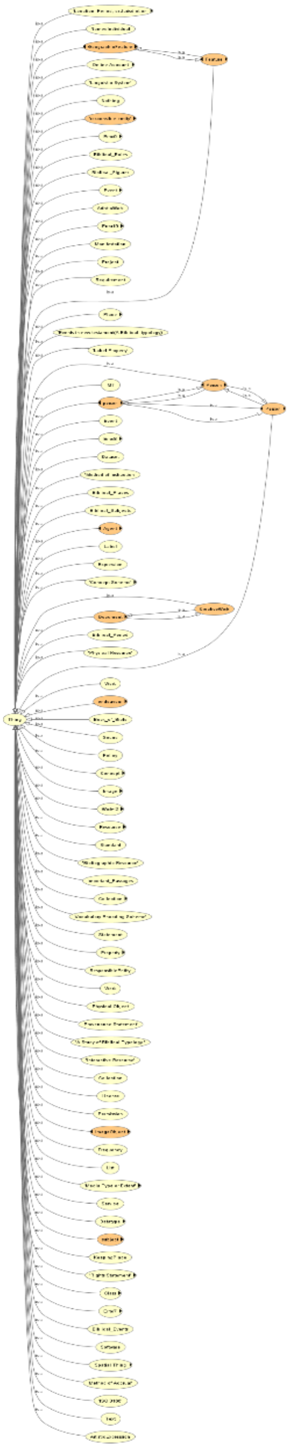
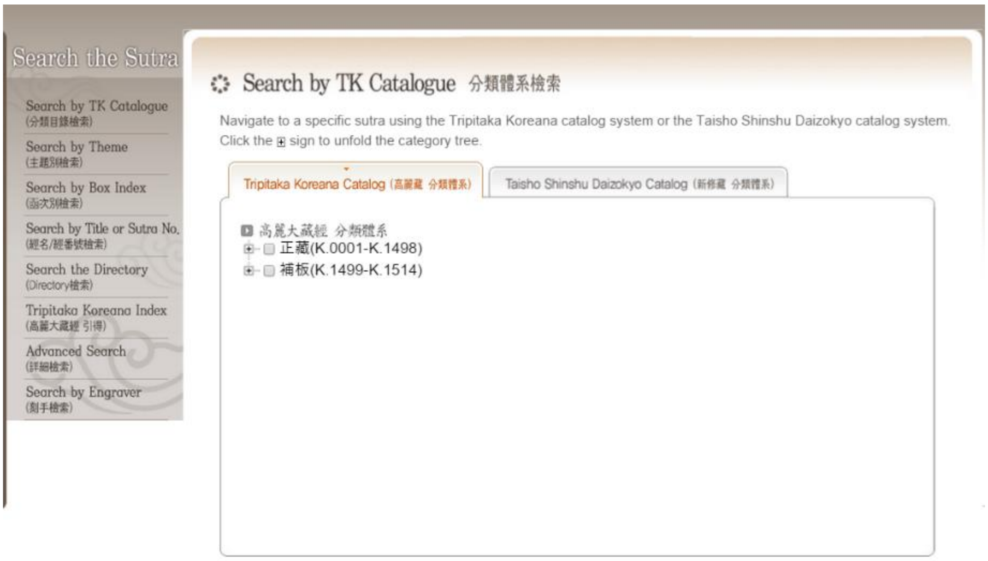


**Figure 6: The Bible ontology graph**

### 2.7.3 Buddhism

Buddhism is a world religion spread mostly across East Asia, started in the 6th century BC on the Indian subcontinent, by Siddhartha Gautama. The sacred text of Buddhism is called the Tipitaka, and is written in the ancient Indian language Pali. This language is very close to the language that the Buddha himself spoke. The Tripitaka is a very large book, and its English translation is split across forty volumes. As a result, some free online applications have been developed to extract information from the Tipitaka, such as the 'Tripitaka Koreana Knowledgebase Project'[40], initiated by the Research Institute of the Tripitaka Koreana. The primary aim of the project is to collect all different types of Tripitaka resources and then make them available as a web application. On this website, the core technique used to retrieve data is the keyword search, which has different features such as searching by a Tripitaka catalogue, theme, index and title, or sutra[41] number. Additionally, this application has a Buddhist dictionary that provides several synonyms and descriptions of words according to context.



Search by TK Catalogue 分類體系檢索

Navigate to a specific sutra using the Tripitaka Koreana catalog system or the Taisho Shinshu Daizokyo catalog system. Click the ⊞ sign to unfold the category tree.

Search the Sutra
Search by TK Catalogue (分類目錄檢索)
Search by Theme (主題別檢索)
Search by Box Index (函次別檢索)
Search by Title or Sutra No. (經名/經番號檢索)
Search the Directory (Directory檢索)
Tripitaka Koreana Index (高麗大藏經 引得)
Advanced Search (詳細檢索)
Search by Engraver (刻手檢索)
Tripitaka Koreana Catalog (高麗藏 分類體系)
Taisho Shinshu Daizokyo Catalog (新修藏 分類體系)
高麗大藏經 分類體系
正藏(K.0001-K.1498)
補板(K.1499-K.1514)

**Figure 7: Tripitaka Koreana Knowledgebase**

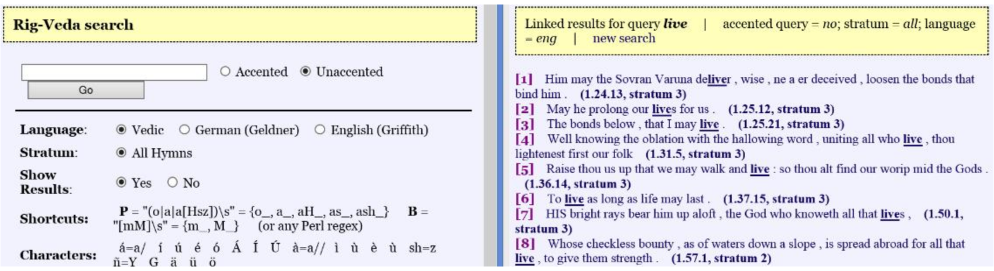
[40] http://kb.sutra.re.kr/ritk_eng/intro/introProject.do

[41] *Sutra is means "discourse". It is a Sanskrit term. an ancient Indo-European language of India, in which the Hindu scriptures and classical Indian epic poems are written and from which many northern Indian (Indic) languages are derived (oxford dictionary).*

### 2.7.4 Hinduism

Hinduism is a major religion of South Asia that developed out of the Vedic religion.  The earliest Hindu religious texts are called Vedas, and they contain hymns, philosophy, and guidance from the Vedic religion. Hindus believe that scholars received these texts from God and have orally conveyed them to the next generations. A few online search tool applications are available for Hindu religious texts; one example is the Rig-Veda Search developed by Kevin Ryan[42].

The Rig-Veda Search[43] enables the user to search the Vedas in one of three languages: English, Vedic, or German. The application uses the keyword search technique based on Perl regular expressions.



**Rig-Veda search**

○ Accented ● Unaccented

**Language**: ● Vedic ○ German (Geldner) ○ English (Griffith)

**Stratum**: ● All Hymns

**Show Results**: ● Yes ○ No

**Shortcuts:** **P** = "(o|a|a[Hsz])\s" = {o_, a_, aH_, as_, ash_} **B** = "[mM]\s" = {m_, M_} (or any Perl regex)

**Characters:** á=a/ í ú é ó Á Í Ú à=a// ì ù è ù sh=z ñ=Y G ä ü ö

Linked results for query ***live*** | accented query = *no*; stratum = *all*; language = *eng* | new search

[1] Him may the Sovran Varuna deliver , wise , ne a er deceived , loosen the bonds that bind him . **(1.24.13, stratum 3)**
[2] May he prolong our lives for us . **(1.25.12, stratum 3)**
[3] The bonds below , that I may live . **(1.25.21, stratum 3)**
[4] Well knowing the oblation with the hallowing word , uniting all who live , thou lightenest first our folk **(1.31.5, stratum 3)**
[5] Raise thou us up that we may walk and live : so thou alt find our worip mid the Gods . **(1.36.14, stratum 3)**
[6] To live as long as life may last . **(1.37.15, stratum 3)**
[7] HIS bright rays bear him up aloft , the God who knoweth all that lives , **(1.50.1, stratum 3)**
[8] Whose checkless bounty , as of waters down a slope , is spread abroad for all that live , to give them strength . **(1.57.1, stratum 2)**

**Figure 8: Rig-Veda Search**

### 2.7.5 Conclusion

In this section, the most popular religious books have been introduced and reviewed in terms of the developed search methods and knowledge representations techniques. This survey aimed at outlining the common information retrieval methods applied to the domain of religious texts and how these techniques could be applied to the Quran.

It was concluded that keyword search methods are mostly commonly used to retrieve knowledge from all religious texts. Additionally, the most keyword search methods used to find information in the religious text are word matching, search by topic, Search by synonyms, morphological search, CRLS, Boolean model, Regular expression, and stemming techniques. All these IR techniques are

[42] *http://www.people.fas.harvard.edu/~kevinryan/*

[43] *http://meluhha.com/newrv/*

used in the current Quran search tools as discussed in chapter 3. Furthermore, advanced IR techniques are recommended to be applied on religious text to increase the performance of current religious search tools.

However, the Quran and the Bible are the sacred texts that have knowledge representation, such as Bible ontology and Quran ontology.

## 2.8 Computational research for the Holy Quran

Research on computational searching of the Quran will be reviewed in Chapter 3, including the explanation of the Holy Quran, the structure of the Holy Quran, Quranic search applications, and existing research on Quranic search techniques and tools.

## 2.9 Conclusion

In this chapter, the different sacred texts of the most popular religious have been introduced and reviewed in terms of the search methods developed for them and knowledge representations techniques. It was explained that keyword search methods are the mostly commonly used methods to retrieve knowledge from all religious texts. However, there are some religious texts for which knowledge representation techniques have also been used, such as Bible ontology and Quran ontology.

# Chapter 3
# Evaluation criteria for computational Quran search

This chapter will review search tools constructed for Information Retrieval from the Holy Quran. Additionally, this chapter evaluates these different search tools against 14 criteria depending on: search features, output features, precision of the retrieved verses, recall database size and types of database contents. Based on this survey, we conclude that most of the existing Quran search tools still cannot solve the problem of ambiguity in the retrieved results because these tools use traditional query analysis and make limited usage of Quran ontologies.

The work presented in this chapter has appeared in the following publications:

Alqahtani, M., & Atwell, E. (2017). Evaluation Criteria for Computational Quran Search. International Journal on Islamic Applications in Computer Science and Technology, 5(1).

Alqahtani MMA; Atwell E (2016) Comparison Criteria for Computational Quranic Search Methods. In 4th International Conference on Islamic Applications in Computer Science and Technologies, Sudan, 20 Dec 2016 - 22 Dec 2016.

## 3.1 Introduction

Both techniques IR and Semantic Search have been applied on the Holy Quran. Many Quran search applications have been built to facilitate the retrieval of knowledge from the Quran. Depending on this study, techniques were used to retrieve information from the Quran can be classified into two types: semantic-based and keyword-based techniques. 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 (Sudeepthi et al., 2012). The majority of Quran search tools employ the keyword search.

The existing Quran semantic search techniques are an ontology-based (A. R. Yauri et al., 2013), a synonyms-set (Shoaib et al., 2009) and a cross-language information retrieval (CLIR) technique (Yunus et al., 2010). The ontology-based technique searches for the concept(s) matching a user query and then returns the verses related to this concept(s). The synonyms-set method produces all synonyms of the query word using WordNet and then finds all Quran verses matching these terms' synonyms. CLIR technique translates words of an input query to another language and then retrieves verses that contain words matching the translated words.

On the other hand, text-based techniques are a keyword-matching, a morphological-based (Al Gharaibeh et al., 2011) and a chatbot technique (Abu Shawar & Atwell, 2004). The keyword-matching method returns verses that contain any query words. The morphological-based method provides a root word search. It generates all other forms of the query word and then finds all Quran verses matching these word forms. The chatbot method selects the most significant or important words from a user query and then returns the Quranic verses containing any words matching the selected words.

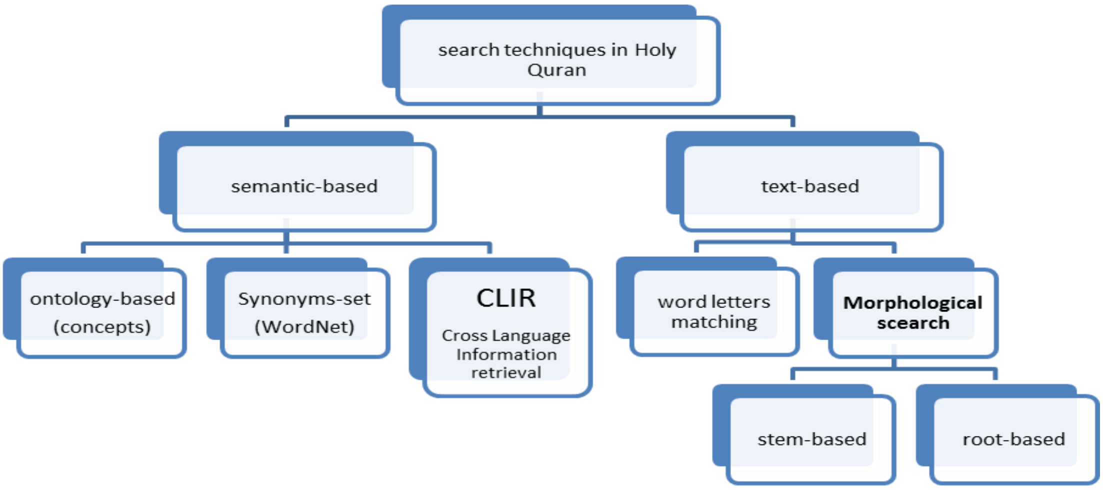


**Figure 9: Classification of Existing Quran Search Techniques**

Several deficiencies exist with the retrieved Quran verses (Aya'at) for a query using the existing keyword search technologies. These problems are: some irrelevant verses are retrieved, some relevant verses are not retrieved or the sequence of retrieved verses is not in the right order (Shoaib et al., 2009). Additionally, the keyword-based techniques have limitations include misunderstanding the exact meaning of input words forming a query and neglecting some theories of information retrieval (Raza et al., 2014).

Moreover, the current Quran semantic search techniques have some limitations concerning finding the requested information. These constraints result in ambiguity in the results because these semantic search tools use one or two incomplete Holy Quran ontologies and ignore the others. Additionally, these ontologies have different scopes and formats that need an alignment and normalization (Alrehaili & Atwell, 2014).

This chapter aims to review and evaluate search tools constructed for the Holy Quran. This objective is achieved by assessing these different search tools against 13 criteria depending on search features, output features, precision and recall of the retrieved verses, database size, and types of database contents.

This chapter is organized as follows. Section 3.2 is literature review containing an analysis of the characteristics of the Holy Quran, Quran search applications, and previous research on Quran search tools. Section 3.3 describes the methodology in terms of evaluation criteria and comparison of different Quran search tools. Finally, Section 3.4 concludes the critical points in this chapter.

## 3.2 Text characteristics of the Holy Quran

Challenging points regarding the text of the Holy Quran exist when applying NLP technologies. First, a concept might be mentioned in different verses. For example, the concept of Hell (النار) [Elnar] is discussed in various chapters and verses, and Allah appears throughout the entire Holy Quran.

An additional Quran feature is that one verse may contain or allude to many topics. For example, verse 40 in Chapter 78 contains seventeen Arabic words, but describing five different concepts.

**إِنَّا أَنذَرْنَاكُمْ عَذَابًا قَرِيبًا يَوْمَ يَنظُرُ الْمَرْءُ مَا قَدَّمَتْ يَدَاهُ وَيَقُولُ الْكَافِرُ يَا لَيْتَنِي**
**كُنتُ تُرَابًا**

[The Quran: Chapter 78, verse 40]

**“Indeed, we have warned you of a near punishment on the Day when a man will observe what his hands have put forth and the disbeliever will say, "Oh, I wish that I were dust!"**

These concepts are: ‘Allah’ has warned ‘Humans’, ‘Allah’ has warned of ‘chastisement’, ‘Man’ will see on ‘the Judgment day’ what his two ‘hands’ did, and ‘Unbelievers’ will say on ‘the Judgment day’ that ‘we wish we were dust’ (Raza et al., 2014). Note also from the above example that this verse does not have the words ‘Allah’ and ‘the Judgment’, but the context reveals what is being said.

Another aspect of the style of the Quran is that one concept is mentioned using many different words, depending on the context. For example, Muhammad is the same as Ahmad, Mudhathir, Muzammil, and messenger of Allah. Another example is that Heaven has different names, such as The Garden and Paradise.

A term may also refer to completely different things, depending on the context: for example, (l-jannat) ‘الْجَنَّة’ refers to the paradise, and the garden. Additionally, two unlike words may have the same letters but have different diacritics. For example, ‘الجنة’ represents three different words: (l-jannat) الْجَنَّة means paradise, (junnat) الجُنَّة means ‘cover’, and (jinnat) ‘الجِنَّة’ means ghosts.

Furthermore, a term might have different names which are not in the synonyms group of this term such as other names of the Paradise as in Figure 10.

The text of the Holy Quran is classical Arabic language, which is different from the modern Arabic language. This difference may cause an incompatibility problem or gap between the query and retrieved verses.

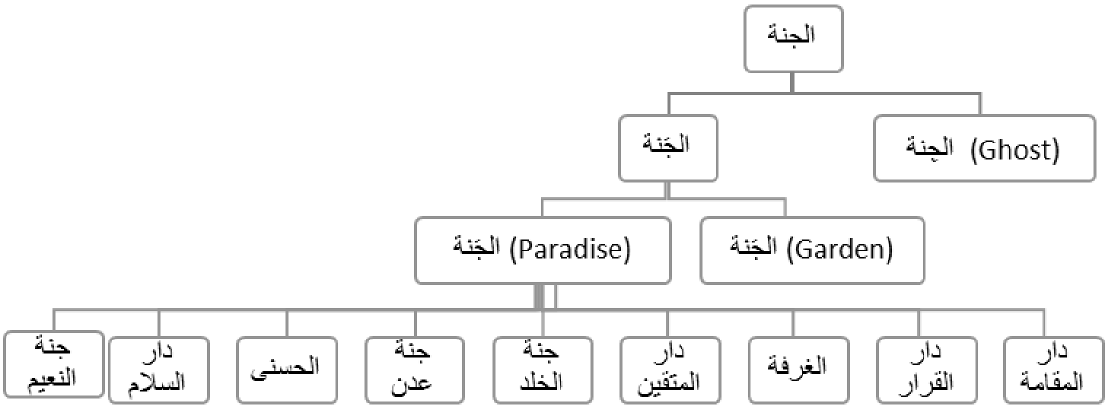


**Figure 10: Ambiguity of Arabic Word in the Quran**

## 3.3 The Quran search applications

Several desktops and Web applications have been developed to retrieve knowledge from the Quran. Many of these applications use keyword search techniques with a few information retrieval methods as outlined in the following.

Khazain-ul-Hidayat[44] and Zakr[45] are free desktop applications that enable the user to read, listen to and search the Quran in many different languages. These applications are mainly designed to be tools for teaching the Quran. A user can search the Quran by querying a word or by entering a verse number. When the user searches for a word, the results will include all verses containing the same query word.

Almonagib Alqurany[46] (المنقب القرآني), Islam web[47] , Tanzil[48] , Quranic Arabic Corpus[49] (QAC) and the Noble Quran[50] are online Web applications that enable users to read, listen to and search the Quran in different languages. Users can search by a chapter number, verse number or word. For instance, a user can search for Chapter 13 and verse 3. Additionally, in the case of searching for a word, these applications will return all verses that have words belonging to the same root of the query word. For example, if the query word is ‘ذكر’’, then the retrieved verses will contain any term of a stem ‘ذكر’’, such as ‘اذكر’’, ‘تذكرة’, ‘الذاكرون’ and ‘ذكرى’.

KSU Quran[51] is a Web application for the study of the Quran and was developed at King Saud University. This Web application represents the Quran in different data forms, such as text and audio in many languages.

The Quran[52] is a project that has been at the forefront of Quran media for several years, with a vision of broadcasting the voice of the Quran. This project was

[44] *http://www.khazainulhidayat.com/*

[45] *http://zekr.org/*

[46] *http://www.holyquran.net/search/sindex.php.*

[47] *http://quran.al-islam.com/*

[48] *http://tanzil.net/*

[49] *http://corpus.quran.com/*

[50] *http://quran.com/*

[51] *http://quran.ksu.edu.sa/*

[52] *http://the Quran.info/*

developed by two former computer science students at the University of Copenhagen in 2007. Many years later, this project has improved by adding more features, such as a search tool, the ability to compare parallel translations of the Quran, and the ability to listen to Quran recitation. Both the websites KSU Quran and the Quran.info provide users with search tools using a word, root or phrase search. The main technique here uses the keyword search, and the retrieved verses are ordered depending on the Quran index.

Semantic Quran[53] is an online search tool application that allows a user to search verses based on concepts. The idea behind this application is that many verses in the Quran relate to certain concepts even though these verses do not have the words commonly associated with the concepts. For instance, Allah says in chapter Ash-Sharh verse (94:5):

[فَإِنَّ مَعَ الْعُسْرِ يُسْرًا][54]

'Indeed, with hardship will be ease'[55]

This verse points to 'hope' and 'patience for believers who are in difficulty', but the words 'hope' or 'patience' do not appear in the verse's words, so these concepts will not be found using a basic keyword search. Therefore, this application used crowdsourcing[56] technique to tag the Quranic verses with concepts. These concepts were created by internet users based on their knowledge. However, not all of the Quranic verses were completely tagged with concepts, when this web application was reviewed by the author.

This Semantic Quran application fails to use various methods to improve the trustworthiness of crowdsourced data including voting system, golden standard, and domain expert users. The voting system employs additional domain experts to judge the crowdsourcing outcome (Barbier, Zafarani, Gao, Fung, & Liu, 2012). The gold standard is the scale that is the best available in a specific condition. Another weakness is that users, who tagged the Quranic verses with concepts, were not experts in the Quran domain which needs specific background

[53] *http://semquran.com/*

[54] *The Quran verse 94:5*

[55] *https://quran.com/94/5?translations=20*

[56] *crowdsourcing is the practice that allows internet users to participate in the tagging of the Quranic verses with concepts*

knowledge. therefore, then using crowdsourcing to tag the Quranic verses with concepts was not the proper approach.

## 3.4 Research on the Quran search tools

Considerable computational research has been carried out on the Quran, including both keyword-based IR and semantic search research.

(Abdelnasser et al., 2014) proposed a new Arabic question-answering system in the domain of the Quran. The system prompts users to enter an Arabic question about the Quran. Then, this system retrieves relevant Quran verses with their Arabic descriptions from ***Ibn Kathir's Tafsir***, a standard commentary textbook on the Quran. This system uses 1,217 Quranic concepts integrated from the Quranic Arabic Corpus Ontology (Dukes, 2013) and Qurany Ontology (N. Abbas et al., 2012; N. H. Abbas, 2009a). It is claimed that the accuracy of the first answer from the retrieved results can reach 65%. This system has three phases for answering a question: question analysis, IR and answer extraction. In question analysis, the 'Morphological Analysis and Disambiguation of Arabic' tool (MADA) (Habash, Rambow, & Roth, 2009) is applied to a user's question to add a part of speech (POS) tag and a stem to each word in this question, and then, all stop words are removed based on their POS tags. Additionally, the remaining words are tagged with Named Entity Recognition (NER) types using the LingPipe tool[57]. For example, the word 'mountain' has a NER type of 'location'. Moreover, this phase uses a support vector machine (SVM) to identify the question type, such as whether it is about a place or a person. In the IR stage, the question is processed via an 'explicit analysis approach' (Gabrilovich & Markovitch, 2007) that enhances a keyword-based text representation with concept-based features in which these features are automatically extracted from the Quran Ontology. After this step, the IR module retrieves related verses from the Quran and their interpretation from Tafsir books. Finally, answer extraction ranks the retrieved answers based on the number of matching words in the answer, the NE type of both the question and answer and the shortest distance between the matched expressions in the retrieved results. This proposed system does not recommend a

[57] *http://alias-i.com/lingpipe/index.html*

solution if the question terms do not match any concepts from the Quran Ontology. Moreover, this tool still does not solve the problem of ambiguity in the results. For example, if a user searches for ‘Elnar : النار’ (fire) the result should tell the user there are two types of ‘النار’: The Hell-fire and normal fire.

(Khan, Saqlain, Shoaib, & Sher, 2013) demonstrated a Quran semantic search method by developing a simple ontology for the animals mentioned in the Quran. The ontology was built using the editor Protégé, and SPARQL is used to answer a query about animals. This paper concludes that the existing Arabic WordNet is not sufficient for finding synonyms for query words in an effort to increase one’s chances of retrieving information from a document. Based on this, they suggested developing Arabic WordNet for Quran words.

(A. R. Yauri, 2014) proposed a semantic search system for retrieving Quran verses based on the enhanced Quranic Arabic Corpus ontology developed by (Dukes, 2013). The system analyses a user question by removing all words except for nouns and verbs. Then, it checks if these words or their synonyms match concepts in the ontology domain. After that, it uses the matched concepts to generate triples in the form of a subject-predicate-object. Finally, it uses SPARQL to answer the user’s query based on generated triples. However, Yauri did not suggest a solution for the ambiguity in the query and the result. Moreover, the proposed search tool is designed for English and Malay translations of the Quran, but not the original Arabic source text.

(Yahya, Abdullah, Azman, & Kadir, 2013) recommended a semantic search for the Quran based on CLIR. They created a bilingual ontology: English and Malay languages. This ontology is also based on the Quranic Arabic Corpus ontology developed by (Dukes, 2013). They did this to experiment on this ontology for two translations of the 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.

(N. Abbas et al., 2012; N. H. Abbas, 2009a) developed a tool called “Qurany” for searching the Quran text in both Arabic and English. In this project, 6,236 HTML pages were created in which each HTML page contains one verse in source Arabic, eight different English translations and the topic(s) of this verse in both Arabic and English. This project’s main idea involves searching the Quran’s eight translations using the keyword search. This method will enhance the precision of the results. Abbas noted that most of the available search tools on the Web use one English translation during the search process, and then return results with average recall and precision values of 54% and 48%, respectively. She showed that the Qurany tool provided an 87% recall value and a 58% precision value. However, this tool uses a basic keyword search when searching for Arabic words. For example, if we search for ( صدق : sidq) the tool will return any word contains letters of صدق’ such as ‘صدقات’.

(Dukes, 2013) developed the Quranic Arabic Corpus, which includes a Web application offering Arabic keyword search and morphological search. These two features enable users to search the Quran by any form of an Arabic word or its POS tag, such as noun, proper noun, and pronoun. This system does not solve the problem of ambiguity in the keyword or the search results. For example, if a user searches for (Elnar: النار) (fire) the result should tell the user there are two types of ‘النار’: The Hell-fire and the normal fire.

## 3.5 Methodology

This section describes the procedure used to assess Quran search tools. This is achieved by selecting common criteria for evaluating different search tools. Then, evaluating the existing Quran search tools that were discussed in section 4 against these criteria. The main aim of this evaluation is to find the key causes of drawbacks and limitations in existing search tools.

### 3.5.1 Criteria for comparing the Quran search methods

In this section, the methodology to evaluate the Quran search tools is mainly based on search algorithms, the accuracy of the result, and Database size. Spiteri and Richard (2013) summarized the most common criteria in 31 articles related to methodologies for the evaluation of search engines. The common measures are search features such as Boolean operators, relevance rankings, recall and precision of results, database size, response time, query type, and database contents. The detailed evaluation criteria for Quran search tools are described in table 3.

## Table 3: Comparison Criteria for Quran Search Methods

| | |
|---|---|
| **1.** Search techniques | **A. Semantic search techniques:** seek to improve search accuracy by analysing a search query in terms of user intent, the contextual meaning of query terms, and search domain. Semantic search aims to overcome the ambiguity of query words and unranked search results. This technique covers the followings:<br>I. Synonym (WordNet): produces all synonyms of the query word using WordNet and then finds all Quranic verses matching these terms' synonyms.<br>II. CLIR: translates words of an input query to another language and then retrieves verses that contain words matching the translation.<br>III. Ontology: searches for the concept(s) matching a user query and then returns the verses related to this concept(s).<br>**B. Keywords Search:**<br>I. Letters are matching: returns verses that contain any query words.<br>II. Morphological search: provides a root word search. |
| **2.** Query analyser | This is a prepossessing stage of a search technique that can be used to analyse a user's query:<br>A. Part of Speech (POS)<br>B. Spelling check<br>C. Named Entities Recognition.<br>D. Stem, and Lemma of query words.<br>E. The root of query words.<br>F. Synonyms of query words.<br>G. None of the above methods. |
| **3.** Quran Ontologies | the existing ontology of the Quran. The Quran ontology is defined as the abstract concepts in the Quran, and the relationships between these concepts. This ontology depends on the knowledge enclosed in traditional sources of the Quran sciences, including the Hadith of the prophet Muhammad (peace be upon him), and the Tafsir (Quranic exegesis). All the following ontology are discussed in chapter 4.<br>A. Arabic Qur'an Corpus Ontology (Dukes, 2013)<br>B. Qurany Ontology [Qur'an topics] (Abbas, 2009, 2013)<br>C. QurAna Ontology: (Sharaf & Atwell, 2012)<br>D. QurSim Ontology: (Sharaf & Atwell, 2012).<br>E. Semantic Quran: (Sherif &Ngonga, 2009)<br>F. Historical concepts: (A. Yauri, Azman, Murad, & Kadir, 2014)<br>G. Animal Ontology in Quran domain (Khan et al., 2013)<br>H. None. |
| **4.** Total number of Quranic Ontologies: used in a search tool based on the list of Quranic ontology in point 3. | |
| **5.** Database size<br>(Quran Datasets) | **A. Quran Arabic text**<br>I. Part of Quran<br>II. All chapters.<br>**B. Translations**<br>I. One English language translation of the original text of the Quran.<br>II. More than one English translation resources. |

| | |
|---|---|
| | **C. Arabic Language Descriptions of the Quran (Tafsirs)**: books written by Islamic scholars to explain and describe the meaning of each verse:<br>I. Tafsir Ibn Kathir.<br>II. Al-Jalaleen.<br>III. Al Qrtubi.<br>IV. Al Muyaser.<br>V. Al-Jaza'iri.<br>VI. More than 5<br>**D. Arabic Quranic Corpus.**<br>**E. Quranic word meanings.**<br>**F. Chapters and verses revelation reasons.** |
| **6.** Number of dataset types: the number of datasets types is involved in search tools (1,2 or 3 ..) | |
| **7.** Query types | input user query type:<br>A. One word.<br>B. Two words.<br>C. Sentences.<br>D. Questions. |
| **8.** Results types | based on the kind of retrieved answer:<br>A. Texts (word, two words, phrase, not a verse, part of the description of verse).<br>B. Verse.<br>C. Combined verses.<br>D. Description of verse.<br>E. Concept(s). |
| **9.** Availability | is this tool available as an application to be used by others?<br>A. Available.<br>B. Not available |
| **10.** Result Ranking | how are the retrieved results ordered?<br>A. Ranked results: ranking the retrieved verses depending on relevance to the conceptual meaning of the query.<br>B. Not ranked: ranking the retrieved verses based on the order of their chapters (chapter 1, 2, ……….114)) |
| *11.* User categories | the target users for this application:<br>A. Public (General use).<br>B. Education<br>C. Islamic Scholar<br>D. Linguistics scholar |
| **12.** Search Domain Coverage | A. All Quran chapters<br>B. Part of the Quran, e.g. one or more specific chapters |
| **13.** Language of the input query | A. Arabic language<br>B. Not-Arabic language |
| **14. Evaluation metrics** | A. Recall<br>B. Precision<br>C. F-measure<br>D. Recall at k<br>E. Precision at k<br>F. Mean Average precision<br>G. None |

### 3.5.2 Comparison of the different existing Quran search tools

The above 14 features described in table 3 are used to compare the different existing Quran search tools that were discussed in section 3.4. Table 4 summarizes the comparison results of the Quran computational search tools which are discussed in section 3.5.

**Table 4: List of evaluated Quran search tools**

| Comparison Criteria / Quranic Search tool | 1. Search techniques | 2. Query analyser | 3. Quranic Ontologies | 4. No. of Ontology | 5. Quran Datasets | 6. Number of dataset types | 7. Query types | 8. Results types | 9. Availability | 10. Result Ranking | 11. Systems Users | 12. Search Domain Coverage | 13. Language of query | 14. Evaluation metrics |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Al-Bayan. (Abdelnasser et al., 2014) | A.III | A, D | A, B | 2 | C.I, C.V | 3 | D | A | B | A | A, C | B | A | E |
| 2. Khan (2013) | A.III | A, D | F | 1 | A.I | 1 | D | B, E | B | B | A | B | B | A,B |
| 3. (Yahya et al., 2013) | A.II, A.III | F | A | 1 | B.II | 2 | A, B, | D | B | B | A | A | B | A,B |
| 4. (Abbas 2009, 2013) | B.I | G | B | 1 | A.II, B.II | 2 | A, B | B, D, E | A | B | A | A | A, B | A,B |
| 5. (Dukes 2012, 2013) | B.I, B.II | F | A | 0 | A.II, D | 2 | A | A, B | A | B | A, B, C, D | A | A, B | G |
| 6. (Shoaib et al., 2009) | A.I | F | H | 0 | A.I, B.I | 2 | A | B | B | B | A, C | B | B | A,B |
| 7. Al Gharaibeh et al., 2011) | A.I, B | D, F | H | 0 | A.II | 1 | B | B | B | B | A | A | A | A, |
| 8. (Yunus et al., 2010) | A.II | F | H | 0 | A.II, B.II | 3 | A, B | B | B | B | A | A | A, B | |
| 9. (Abu Shawar & Atwell, 2004) | B.I | G | H | 0 | A.II | 1 | A, B, C | B | B | B | A | A | A | C |
| 10. Quran.com | B.I, B.II | E | H | 0 | A.II, B.II, C.II | 3 | A, B, C | B, D | A | B | A, C | B | A, B | G |
| 11. Tanzil.net | B.I, B.II | F | H | 0 | A.II, B.II, C.II, C.IV | 4 | A, B | B | A | B | A, C | B | A | G |

## 3.6 Conclusion

This chapter summarized the search techniques used in the existing Quranic search tools such as desktop applications and online applications. Additionally, this chapter reviewed the previous studies on the Quranic search tools. After that, the Quran search tools were evaluated against 14 criteria. The outcome evaluation of the search tools, based on table 4, shows several deficiencies in the current Quranic search models.

The first weakness is that, most semantic search tools used only a unique source or part of the existing Quran ontologies. For example, in table 4 the semantic search tools 1, 2, and 3 used only 2 Quranic ontologies which are not cover all aspects in the Quran. Therefore, this will affect the accuracies of the retrieved results. To overcome this weakness, the existing Quranic ontologies need to be integrated and expanded by including more Quranic resources. In this study, the Quranic Ontologies will be developed using the common semantic web standards as described in chapter 4.

The second limitation, these search tools prompted users to search by only at max two query types as can be seen in table 4. Therefore, a new Quranic search tool should allow users to search by any type of queries such as concepts, phrases, sentences, or questions. This limitation could be solved by adding a natural language query analyser to the search tools to handle any type of queries including words, phrases and questions (this will be dissuaded in chapter 6).

The third drawback, these tools fails to use advanced methods to analyse the query texts by applying NLP techniques, including parsing, a question classification and POS tagging. To solve this issue, the query analyser is recommended to apply advanced NLP techniques such as POS tagging, and the question classification.

The fourth restriction in these tools is that absence of well-formatted Arabic Named Entities lists specialized for the Quran text, such as prophets' names, Allah's names, animals, times and religion. Consequently, the search tools fail to predict answer type of an input question. Section 5.3.6 will tackle this issue by building a new Arabic Quranic question classifier using a list of Quranic question classes.

The final limitation is that, these search models fail to use the appropriate IR evaluation metrics for ranked results to measure the performance of the search tools. For example, table 4 shows that all search tools were evaluated by IR systems metrics of unranked retrieved verses. This study will develop a suitable testing dataset to examine any Quranic search tool as explained in section 5.3. After that, the study will examine the proposed Arabic semantic search model against this  dataset. Finally, the testing results of the model will be  evaluated using the IR systems metrics for ranked results as described in section 6.3.1.

# Chapter 4
# Developing bilingual Arabic-English ontologies of the Quran

This chapter will review and evaluate current Quranic ontologies. Then it will present several stages of developing the new Arabic-English Quran ontology from different datasets related to Al Quran.

The work presented in this chapter has appeared in the following publications:

Alqahtani MMA; Atwell E (2016) Aligning and Merging Ontology in the Quran Domain. In the 9th Saudi Students conference in the UK, Birmingham, 13 – 14 Feb 2016.

Alqahtani, M. M., & Atwell, E. (2018, March). Developing Bilingual Arabic-English Ontologies of the Quran. In *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)* (pp. 96-101). IEEE.

## 4.1 Introduction

Many research studies have been built on ontologies to facilitate the retrieval of knowledge from the Quran. The primary goal of developing Quranic Ontologies is to enrich the raw Arabic Quran text with Islamic semantic tags. The term ontology is defined as an explicit specification of concepts, attributes and relations in a domain (Gruber, 2009). Common components of ontologies include: classes that are concepts, attributes, relations, function terms, restrictions, and axioms. The concepts are entities of interest in a particular domain. These concepts are structured into a taxonomy tree or un-taxonomy tree. Each tree node represents a concept that is a specialization of its ancestor. The concept is related to a set of instances. Additionally, the concept has a set of attributes. The relations are ways in which concepts and instances can be linked to each other. More details about the ontology were discussed in section 2.5.

In this study, the Quran ontology is an essential component of the Arabic Quranic semantic search model. Consequently, before developing a new Quran ontology, the existing Quran annotated datasets and ontologies must be reviewed and evaluated. The ontology evaluation is crucial for ontology development as it is an essential stage in the common methodologies of ontology development( previously discussed in section 2.5.2).

An ontology can be evaluated against several criteria, such as the coverage of a particular domain and the size of the ontology. Additionally, ontologies can also be assessed in terms of the specific use cases, scenarios, requirements, applications and scope. This evaluation includes the consistency and completeness of the ontology and the representative modelling language. Moreover, the assessment of an ontology covers the feasibility of the ontology's alignment with other ontologies and improvements (Obrst, Ashpole, Ceusters, Mani, & Smith, 2007). Ahmad and Alrehaili (Ahmad et al., 2013; Alrehaili & Atwell, 2014) compared existing Qur'anic ontologies against 9 criteria. Examples of these criteria are: number of concepts, availability, relation type, verification methods, coverage area, maturity level, and underlying format. These surveys concluded that these ontologies have unclear semantic annotation format, and no validation methods. Therefore, some of the Quran ontologies need normalization and alignment (described previously in section 2.5.3).

The main benefit of merging Qur'anic Ontologies is to pioneer research linking the raw Arabic Quran text with Islamic ontology. Additionally, this combined ontology could help readers to understand the Quran by exploring the hierarchy of the Quran concepts.

Moreover, aligning the Quranic ontologies will increase the coverage of the Quranic ontology in many different knowledges. Furthermore, the alignment will enhance the knowledge extraction from the Quran by increasing the relationships between the Quran concepts. Moreover, alignment of the similar Quranic ontologies will reduce the redundancy of concepts and improve the data quality.

The key challenge in the Quranic ontologies is that these ontologies did not follow the ontology development standards. This caused different files' structure and format of these ontologies. additionally, different spelling of concepts in Arabic and English languages such as 'Moses' and 'Musa', 'Mohammad' and 'Muhammad'. Consequently, three stages are used to align the Quranic ontologies: normalization, terminological approach and structural approach (described previously in section 2.5.3). In the normalization process, all ontologies are reformatted to have the same file format.

The purpose of this study is to review and evaluate most of the Quranic ontologies to overcome their limitations in the development of a new Quranic ontology. the new ontology will use suitable existing Quranic ontology combined with other valuable Quranic datasets. This objective is achieved by assessing 13 different ontologies against 14 criteria.

This chapter is organized as follows. Section 4.2 is a literature review of the Quranic ontologies. Section 4.3 is the methodology of developing and combing the new Quranic Ontology. Finally, Section 4.4 concludes the critical points in this work.

## 4.2 Research conducted on the ontology of the Holy Quran

Sherif and Ngonga Ngomo (2009) developed a Semantic Quran dataset in an RDF format representing translations of the Quran in 42 different languages. This dataset was built by merging data from two different semi-structured sources: the Tanzil project[58] and the Quranic Arabic Corpus[59]. An ontology of the Semantic Quran was constructed to demonstrate various multilingual data from Quranic sources with a hierarchical structure, which is a chapter, a verse, a word and a lexical item. This ontology has 7,718 links to DBpedia[60], 18,655 links to Wiktionary and 15,741,399 triples. The scope of this ontology is translation and structure of the Quran. For example, each verse has translations in 42 different languages. However, the main weakness of this ontology is the failure to provide parallel translations in different languages for each Arabic word in the Quran. For example, the word ‘"m~iSora" ‘مصر’ ‘Eygpt’ in the location(chapter 12- verse 21- word 5) has quran12-21-5-ar ‘"m~iSora" and quran12-21-5-en ‘from’. Another weakness is that this ontology not include other aspects in the Quran.

Khan (2013) developed an ontology for the Quran in the scope of the animals found in the Quran. This ontology was constructed by the editor Protégé, and SPARQL was then used to search through this ontology. The ontology provides 167 links to animals in the Quran based on information found in “***Al-Hayawany Fi el Quran Al-Kareem***” book (El-Naggar, 2006). However, this ontology fails to cover all knowledge in the Quran and is not available to use. Additionally,

[58] *http://tanzil.net/*

[59] *http://corpus.quran.com/*

[60] *http://dbpedia.org*

animal concepts are included in the ontology of the Quran created by Dukes (2013).

Yauri (2013) rebuilt the existing ontology created by Dukes (2013) using the Protégée tool and Manchester OWL. He increased the number of relationships from 350 to about 650 based on the Quran, the Hadith and some online Islamic resources. This ontology scope is to cover most of the subjects mentioned in the Quran, such as food, people, religions and life. However, this ontology does not cover all knowledge in the Quran and is not available to use. An additional drawback is that this ontology was not evaluated by an Islamic scholar.

Yahya (2013) created bilingual ontology featuring the English and Malay languages. This ontology is based on the ontology developed by Dukes (Dukes 2013). They did this to experiment on this ontology for two the Quran translations. In the Malay translation, 5,999 verses are assigned to concepts, and 237 verses are unrelated 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. Nevertheless, this ontology is a duplication of the Quran ontology created by Dukes (2013).

Abbas (2009) developed nearly 1,100 Quranic concrete and abstract concepts linked to all verses of the Quran. She used existing Quranic topics from the Islamic scholarly book called ***Mushaf Al Tajweed*** (Dar al-Maarifah, 1999) . These concepts in the index have an aggregate relationship; the hierarchy of concepts is non-reflexive, non-symmetric and transitive. However, Abbas does not follow any of ontology development methodology to build this annotated dataset. However, this dataset could be a valuable resource to construct the new Quran ontology.

Dukes (2013) extracted 300 concepts and 350 relations from the Quran. The relationship types connecting concepts using predicate logic are Part-of and IS-A. The ontology is based on a famous Quran description book called "***Tafsir Ibn Kathir***" (Abdul-Rahman, 2009). However, this ontology does not cover all concepts in the Quranic verses. Therefore, this ontology could be part of the new Quranic ontology.

Ta'a, Abdullah, Ali, & Ahmad (2014) created a Quranic ontology based on themes mentioned in "***Syammil Quran Miracle the Reference***" ((Indonesia) &

Agama, 2010) . This ontology was evaluated by some experts in the Quran knowledge. This ontology was built using protégé tool in English-Malay languages. However, this ontology is similar to Qurany ontology in which these ontologies have the same topics that were represented in different natural languages. Additionally, the developers of this ontology fail to deploy this ontology to The Linked Open Data Cloud. As consequent, this ontology is not available to download or used.

Muhammad (2012) developed an ontology for the Quran in the scope of pronoun antecedents. This ontology consists of 1,050 concepts and more than 2,700 relations. In addition, 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 (Sharaf & Atwell, 2006).

Aldhubayi and Noorhan Abbas (2012) unified three different Quranic datasets created by former researchers at the University of Leeds. These datasets are the Quran Arabic Corpus (Dukes, 2013), the Quran annotated with Pronominal Anaphor [QurAna] (Sharaf & Atwell, 2012) and the Qurany project (N. H. Abbas, 2009b). These datasets are merged in one XML file, and then the file is uploaded in the Sketch Engine tool as a unified Quranic corpus.

Abdelnasser ( 2014) developed 1,217 leaf concepts linked to the Quranic verses. These annotated datasets were integrated from the concepts in the Quranic Arabic Corpus Ontology [QCO] (Dukes, 2013) and the Qurany topics (Abbas, 2009). In this dataset, each verse of the Quran is connected to at least one leaf concept. However, when QCO and Quranic topics were merged and manipulated, 621 verses did not link to any concepts. This was solved by connecting these un-annotated verses to their relevant concepts. However, this work uses the Quran concepts as a tag set for the Quranic verses without relationships between these concepts.

Hakkoum & Raghay (2016) developed a new Quran ontology by combining the Quran Arabic Corpus (Dukes, 2013), the Quran annotated with Pronominal Anaphor [QurAna] (Sharaf & Atwell, 2012), part of Quranic Arabic Corpus Ontology [QCO] (Dukes, 2013), and the Qurany project (N. H. Abbas, 2009b). These datasets are merged in one OWL file. However, this work fails to include important datasets such as the Quran words meaning, chapters' names and verses' names.

After reviewing the above ontologies, 13 evaluation criteria are designed to assess these 12 different ontologies as explained in section 4.3.1 .

## 4.3 Methodology of developing Quran ontologies

There is no standard methodology to develop a new ontology for a specific domain. However, several outstanding methodologies exist for constructing the new ontology. The methodology of developing the Quranic ontologies follows the life cycle stages shared by the most common practises for ontology development. These phases are: specification, conceptualisation, formalisation, implementation, evaluation, and documentation(described previously in section 2.5.2).

Figure 11 explains the procedure of developing a new Quranic ontology. this procedure went through sequential several stages as follows: Evaluation the Quran ontologies, formatting and normalization of datasets, and developing ontology.

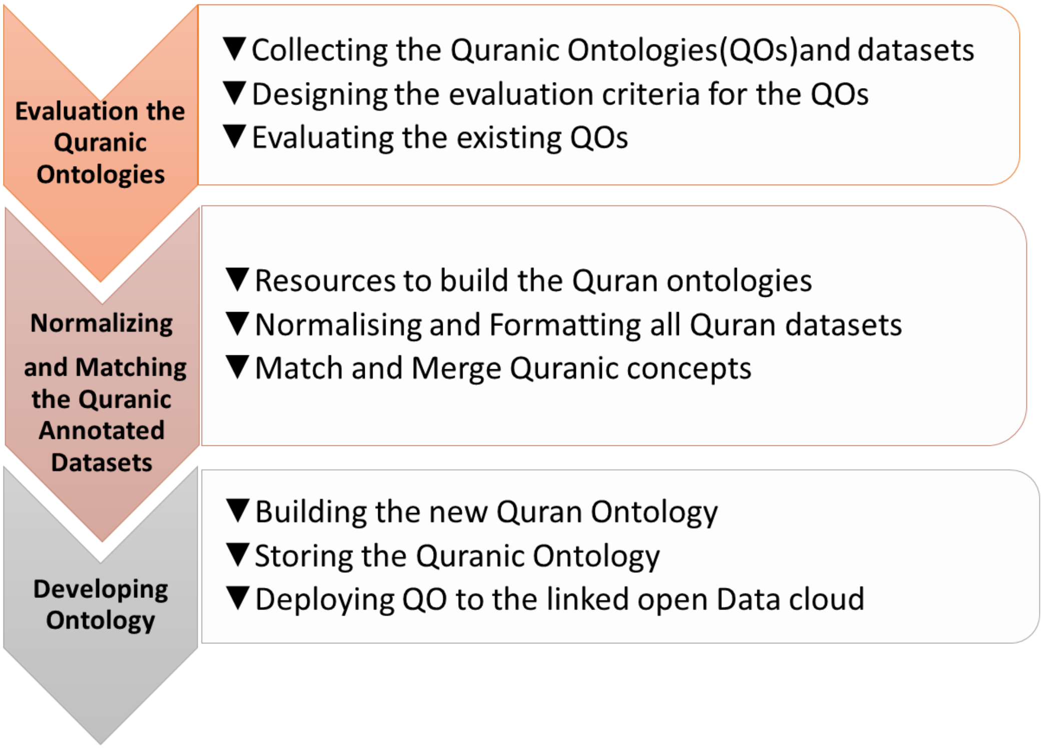

Evaluation the Quranic Ontologies
▼Collecting the Quranic Ontologies(QOs)and datasets
▼Designing the evaluation criteria for the QOs
▼Evaluating the existing QOs
Normalizing and Matching the Quranic Annotated Datasets
▼Resources to build the Quran ontologies
▼Normalising and Formatting all Quran datasets
▼Match and Merge Quranic concepts
Developing Ontology
▼Building the new Quran Ontology
▼Storing the Quranic Ontology
▼Deploying QO to the linked open Data cloud

**Figure 11: Methodology of developing the new Quranic Ontologies**

### 4.3.1 Evaluation the Quranic Ontologies

The evaluation stage contains three steps as in figure 11: collecting the Quranic ontologies and datasets, designing the evaluation criteria for the QOs and evaluating the existing QOs. This stage aimed at assessing the existing ontologies to find which ontologies are valid to be reused or reconstructed in this study.

#### 4.3.1.1Collecting the Quranic ontologies and annotated datasets

In this step, 13 of the Quranic ontologies and the annotated Quranic datasets were collected and reviewed. Nevertheless, 12 out of 13 Quranic ontologies are not available in the linked open data cloud[61]. Additionally, few of Quranic ontologies and another annotated Quranic dataset were downloaded from different locations. These ontologies were previously discussed in section 4.2 .

[61] *https://lod-cloud.net/*

#### 4.3.1.2 Evaluation criteria of the existing Quranic ontologies

The key advantages of evaluating ontology leads to develop a better ontology are: increasing the availability and reusability of ontologies, and ease maintenance of collaboratively created knowledge bases. Nevertheless, bad quality ontology adversely affects the ontology readability such as having vocabulary consists of errors. Moreover, reasoners might be incapable of inferring the right answers in the case of inconsistent semantics (Vrandeči, 2010).

This stage aims at designing criteria to review and evaluate most of the ontologies that are constructed for the Holy Quran in order to reuse the good quality ontologies. An ontology can be evaluated against several criteria. For example, ontologies can be assessed in terms of the specific use cases, scenarios, requirements, applications, triples size, and scope. Additionally, this evaluation includes the consistency and completeness of the ontology and the representation modelling language. Moreover, assessment of ontology covers the feasibly of ontology alignment with other ontology and improvement (Obrst et al., 2007).

In this section, the evaluation of existing Quran Ontologies uses fourteen criteria. These measures are designed to fulfil the aim of this study by using the best common measures of the ontology evaluation including (Ahmad et al., 2013; Allemang & Hendler, 2008; Alobaid, n.d.; Alrehaili & Atwell, 2014; Obrst et al., 2007; Vrandeči, 2010) . The fourteen criteria are as follows:

1. Scope:

    a) Morphological

    b) Translation

    c) Quran topics

    d) Antecedent pronouns

    e) Animals

    f) Time

    g) Subjects

    h) History

    i) Prayer (Salaht)

j) Women

k) Similarity between verses

2. Types of relationships between concepts:

a) Taxonomy or Hierarchy: such as (is_a or part_of, sub_class).

b) Un-taxonomy: uses a verb to describe the relationship between two concepts.

3. Relationship numbers (triples)

4. Number of concepts

5. Semantic Ontology formats

a) Not applicable ( Text)

b) RDF

c) OWL

6. Ontology representation language:

a) Arabic language

b) English language

c) Malay language

d) Dutch language

e) More than 4 languages

7. Source file of the ontology:

a) Available to use

b) Not available to use

8. Validation techniques: methods of validating the ontology:

a) By domain experts: an Islamic scholar

b) Depending on existing Islamic resources such as books.

c) None

9. Coverage Domain of concepts:

a) Cover all Quranic verses

b) Cover almost all Quranic verses

c) Cover half of the Quranic verses

d) Some verses

10. Is dependent on another ontology (dependency): this means that a new ontology is built based on a previous ontology:

a) No.

b) Yes.

11. Is used by another ontology (usage): this means that a new ontology is built based on a previous ontology.

a) Yes.

b) No.

12. Published on Linked Open Data could (availability):

a) Yes.

b) No.

13. Linked to another linked data: upper ontology, such as friend-of-friend ontology:

a) Yes.

b) No.

14. Is ontology used in application:

a) Yes

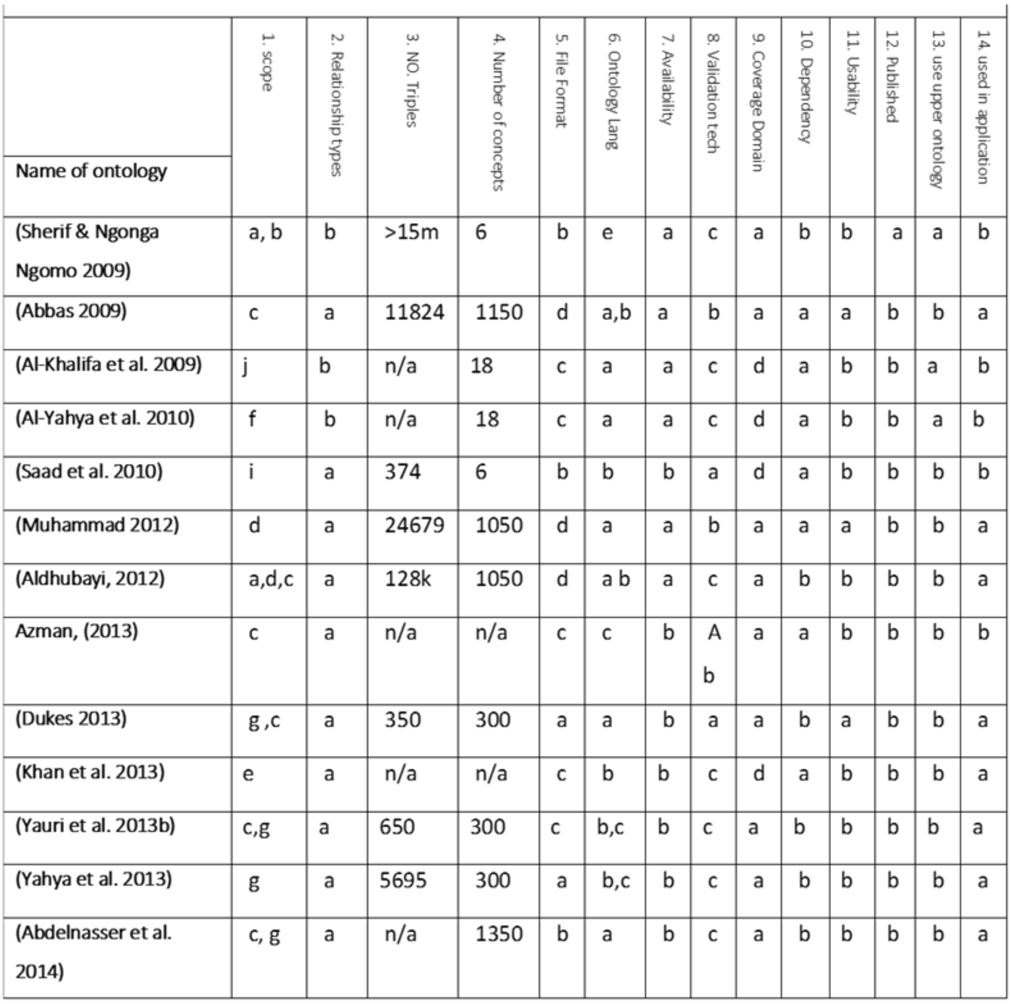
b) No

#### 4.3.1.3 Evaluation of the existing Quran ontologies

Depending on the result of the evaluation of the Quran ontologies in table 5, some deficiencies are found in most of these ontologies. For example, some ontologies fail to be evaluated by an Islamic scholar or tested by an application. Moreover, all of these ontologies do not tag all Quranic verses with semantic tag(s). Furthermore, these ontologies were built in different structures and file formats such as CVS, XML, RDF, OWL or text. Additionally, some ontologies fail to be presented by both Arabic and English languages. Moreover, these different datasets have some similarity in concepts (Overlapped). Additionally, the most of ontologies are part or depending on Quranic topics QT (Abbas 2009), Arabic Quran Corpus AQC (Dukes 2013), Ontology of Quranic Concepts OQC or QurAna (Muhammad 2012).

It is necessary to develop a new Quranic ontology to overcome the limitations of the existing Quranic ontologies . This could be achieved by developing ontology that follows semantic web standards and covers all aspects in the Quan domain. This ontology will be easy to maintain, to reuse and to expand.

**Table 5: Evaluation of 13 Quran Ontology against 14 criteria**



| Name of ontology | 1. scope | 2. Relationship types | 3. NO. Triples | 4. Number of concepts | 5. File Format | 6. Ontology Lang | 7. Availability | 8. Validation tech | 9. Coverage Domain | 10. Dependency | 11. Usability | 12. Published | 13. use upper ontology | 14. used in application |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (Sherif & Ngonga Ngomo 2009) | a, b | b | >15m | 6 | b | e | a | c | a | b | b | a | a | b |
| (Abbas 2009) | c | a | 11824 | 1150 | d | a,b | a | b | a | a | a | b | b | a |
| (Al-Khalifa et al. 2009) | j | b | n/a | 18 | c | a | a | c | d | a | b | b | a | b |
| (Al-Yahya et al. 2010) | f | b | n/a | 18 | c | a | a | c | d | a | b | b | a | b |
| (Saad et al. 2010) | i | a | 374 | 6 | b | b | b | a | d | a | b | b | b | b |
| (Muhammad 2012) | d | a | 24679 | 1050 | d | a | a | b | a | a | a | b | b | a |
| (Aldhubayi, 2012) | a,d,c | a | 128k | 1050 | d | a b | a | c | a | b | b | b | b | a |
| Azman, (2013) | c | a | n/a | n/a | c | c | b | A b | a | a | b | b | b | b |
| (Dukes 2013) | g ,c | a | 350 | 300 | a | a | b | a | a | b | a | b | b | a |
| (Khan et al. 2013) | e | a | n/a | n/a | c | b | b | c | d | a | b | b | b | a |
| (Yauri et al. 2013b) | c,g | a | 650 | 300 | c | b,c | b | c | a | b | b | b | b | a |
| (Yahya et al. 2013) | g | a | 5695 | 300 | a | b,c | b | c | a | b | b | b | b | a |
| (Abdelnasser et al. 2014) | c, g | a | n/a | 1350 | b | a | b | c | a | b | b | b | b | a |

### 4.3.2 Normalizing and Matching the Quranic Annotated Datasets

#### 4.3.2.1 Resources to build the Quran ontologies

twelve deferent resources are used to build the Quran ontology as follows:

- **Quran metadata**[62]: consists of various information about the structure of the Quran text; mainly about chapters, verses and the divisions of the Quran. The chapter (surah) metadata items are number of verses, revelation order, chapter index, number of rukus[63], and chapter name in Arabic, English language and English Transliteration of the Arabic name. the divisions include juz[64] (division), hizb (group), manzil[65] (station, stage), ruku (section), and page (page number in Medina Mushaf). Verse metadata items are: verse index, chapter index, verse order in a chapter, and has sajdah[66].
- **Quran simple Arabic text** tanzil.com: this data set consists of the plain Arabic text for each Quranic verse. This represents the Arabic label for each owl:individual of the class:verse. For example: the Quran verse 1:4 is represented as:

```
<owl:NamedIndividual rdf:about="&Resource;quran1-4">
      <rdf:type rdf:resource="&Resource;Verse"/>
      <rdfs:label xml:lang="ar">مالك يوم الدين</rdfs:label>
</owl:NamedIndividual>
```

- **Othmani Arabic text with diacritical marks** from tanzil.com: this data set consists of the Arabic text with supplementary diacritics for each Quranic verse. This represents the owl:DatatypeProperty ‘displayedText’ for each owl:individual of the class verse. For example: the Quran verse 1:4 is represented as:

*62 http://tanzil.net/docs/quran_metadata*

*63 https://en.wikipedia.org/wiki/Ruku*

*64 https://en.wikipedia.org/wiki/Juz%27*

*65 https://en.wikipedia.org/wiki/Manzil*

*66 https://en.wikipedia.org/wiki/Sujud#Sajdah_of_recitation_/_Tilawah*

```
<owl:NamedIndividual rdf:about="&Resource;quran1-4">
      <rdf:type rdf:resource="&Resource;Verse"/>
      <displayedText xml:lang="ar">مَالِكِ يَوْمِ الدِّينِ</displayedText>

</owl:NamedIndividual>
```

- **Quran English translation text** from tanzil.com: this contains English translation for each Quranic verse. This represents the English label for each owl:individual of the class:verse. For example: the Quran verse 1:4 is represented as:

```
<owl:NamedIndividual rdf:about="&Resource;quran1-4">
      <rdf:type rdf:resource="&Resource;Verse"/>
      <rdfs:label xml:lang="en">Master of the Day of Judgment.</rdfs:label>

</owl:NamedIndividual>
```

- **Tafseer Al-Muassar**[67] **:** this dataset contains a brief description for each Quranic verse in Modern Standard Arabic (MSA). This represents the owl:DatatypeProperty 'descByMuyasser' for each owl:individual of the class verse. For example: the Quran verse 1:4 is represented as:

```
<owl:NamedIndividual rdf:about="&Resource;quran1-4">
<rdf:type rdf:resource="&Resource;Verse"/>
<descByMuyasser xml:lang="ar">
                وهو سبحانه وحده مالك يوم القيامة، وهو يوم الجزاء على الأعمال.
        وفي قراءة المسلم لهذه الآية في كل ركعة من صلواته تذكير له باليوم الآخر،
        وحثٌّ له على الاستعداد بالعمل الصالح، والكف عن المعاصي والسيئات.
</descByMuyasser>
    </owl:NamedIndividual>
```

- **Tafseer Al-Jalalien**[68]: this is the description of the Quran verses in Arabic language. Additionally, this dataset explains Quranic phrases in some

[67] *http://quran.qurancomplex.gov.sa/Quran/tafseer/Tafseer.asp?t=MOYASAR&TabID=3&SubItemID=5&l=arb&SecOrder=3&SubSecOrder=5*

[68] *http://tanzil.net/trans/ar.jalalayn*

verses. This represents the owl:DatatypeProperty 'descByMuyasser' for each owl:individual of the class verse. For example: the Quran verse 1:4 is represented as:

```
<owl:NamedIndividual rdf:about="&Resource;quran1-4">
<rdf:type rdf:resource="&Resource;Verse"/>
<descByJalalayn xml:lang="ar">
أي الجزاء وهو يوم القيامة، وخص بالذكر لأنه لا ملك ظاهرًا فيه لأحد إلا الله تعالى
بدليل «لمن الملك اليوم؟ لله» ومن قرأ مالك فمعناه الأمر كله في يوم القيامة
أو هو موصوف بذلك دائمًا «كغافر الذنب» فصح وقوعه صفة لمعرفة.
</descByJalalayn>
</owl:NamedIndividual>
```

- **Quran word meaning** from ***Mushaf Al Tajweed*** book (Dar al-Maarifah, 1999): this dataset explains some Quranic words and phrases in some verses. This represents the word owl:DatatypeProperty 'wordMeaning' for each owl:individual of the class verse. For example: the Quran verse 1:4 is represented as:

```
<owl:NamedIndividual rdf:about="&Resource;word10-13-3">
      <rdf:type rdf:resource="&Resource;Word"/>
      <rdfs:label xml:lang="ar">القرون</rdfs:label>
      <wordMeaning>الأمم كقوم نوح و عاد و ثمود</wordMeaning>
</owl:NamedIndividual>
```

- **Quran chapter alternative name and some verses names dataset:** these names the author extracted them manually form "***names of Alqauran and its surah names and verses names***" (Adam 2009). This dataset contains the Quranic chapters alternative names and some verses names.

- **Arabic Quran corpus** (AQC)[69]: This dataset consists of the Arabic morphology for each word in the Quan. This dataset will represent the Quran words with their datatype properties such as a word_location in the Quran , a word_translation, a word_lemma, a displayed_Text, and a

[69] *http://corpus.quran.com/*

word_Translitration. For example, the word ‘the generations’ (الــقــرون) is represented as:

```
<owl:NamedIndividual rdf:about="&Resource;word10-13-3">
      <rdf:type rdf:resource="&Resource;Word"/>
      <rdfs:label xml:lang="ar">القرون</rdfs:label>
      <chapterIndex rdf:datatype="&xsd;nonNegativeInteger">10</chapterIndex>
      <verseIndex rdf:datatype="&xsd;nonNegativeInteger">13</verseIndex>
      <wordIndex rdf:datatype="&xsd;nonNegativeInteger">3</wordIndex>
      <wordTranslation rdf:datatype="&rdfs;Literal">the generations</wordTranslation>
      <wordTranslitration>l-qurūna</wordTranslitration>
      <wordMeaning>الأمم كقوم نوح و عاد و ثمود</wordMeaning>
      <wordLemma>قرن</wordLemma>
      <displayedText xml:lang="ar">ٱلْقُرُونَ</displayedText>
      <IsPartOf rdf:resource="&Resource;quran10-13"/>
   </owl:NamedIndividual>
```

- **Quran ontology corpus** QOC[70]: this dataset contains 300 abstract concepts with 350 relationships between them. This dataset will represent the Quranic classes belonging to the category class as shown in figure 12.

- **Quran topics** from ***Mushaf Al Tajweed*** book (Dar al-Maarifah, 1999)

- **QurAna**[71]: this is a corpus of the **Qur**an annotated with Pronominal **Ana**phora. This corpus is developed from the Quran text in which each personal pronoun is tagged with its antecedence concept.

#### 4.3.2.2 Normalising and Formatting all Quran datasets

All the datasets (previously described in section 4.3.21) were collected as XML, plain text or CVS files. Consequently, the different files formats of these datasets were converted into a unique file format as CVS files. After that, all these CVS files were stored in a database. Examples from each dataset are shown in Appendix B.

[70] *http://corpus.quran.com/ontology.jsp*

[71] *http://textminingthequran.com/*

#### 4.3.2.3 Match and Merge Quranic concepts form the Quran Datasets

The current ontologies: OQC, QT, and QurAna are selected to be included in the resources of the new Quran ontology. This is because: they contain Quranic categories, topics and concepts which are connected to the Quran verses. In addition, all these ontologies are represented in both Arabic and English languages, are used by another Quranic ontology, cover different aspects of knowledge, and are available to use. Even though, any of these ontologies does not cover all verses of the Quran as can be seen in table 6, the combination of these ontologies covers nearly 100 per cent of the Quranic verses.

Therefore, similar concepts between these datasets should be identified and matched to reduce the redundancy and increase the data quality. The similar concepts in OQC, QT and QurAna are matched by two methods: an exact match of concepts and a Simple Fuzzy String Similarity.

**Table 6: The semantic tags coverage of the Quran by selected ontologies**

| Ontology | # of Tagged Chapters | # of Tagged verses | Coverage % |
|---|---|---|---|
| OQC | 97 | 1343 | 21.54 |
| QurAna | 112 | 5537 | 88.79 |
| QT | 114 | 5561 | 89.18 |
| All | 114 | 6202 | 99.45 |

The exact match method matches the exact concepts between different datasets. This method was applied three times on the three datasets.

The first experiment compared the concepts’ Arabic labels of all datasets to find the exact match between concepts’ Arabic labels as shown in table 7.

The second experiment compared the concepts’ English labels of all datasets to find the exact match between concepts’ English labels as shown in table 8. The last experiment compared the both concepts’ Arabic and English labels of all datasets to find the exact match between concepts in both Arabic and English labels as shown in table 9.

**Table 7: Arabic concepts matching in the Quran datasets**

| OQC | QT | QurAna | # of matches |
|---|---|---|---|
| ✓ | ✓ | | 39 |
| | ✓ | ✓ | 68 |
| ✓ | | ✓ | 67 |
| ✓ | ✓ | ✓ | 21 |

**Table 8: Exact match of English Quranic concepts match**

| OQC | QT | QurAna | # of matches |
|---|---|---|---|
| ✓ | ✓ | | 12 |
| | ✓ | ✓ | 42 |
| ✓ | | ✓ | 46 |
| ✓ | ✓ | ✓ | 6 |

**Table 9: Both Arabic and English concepts matching in the Quran datasets**

| OQC | QT | QurAna | # of matches |
|---|---|---|---|
| ✓ | ✓ | | 8 |
| | ✓ | ✓ | 23 |
| ✓ | | ✓ | 26 |
| ✓ | ✓ | ✓ | 3 |

The another method was used to find the similarities between concepts is the Simple Fuzzy String algorithm in which the pair of ontology datasets are compared each time (Stragand, 2011). The simple Fuzzy string measures the similarity between two strings using Dice's coefficient which is obtained by the following formula:

$$similarity = \frac{2(a \cap b)}{(a + b)}$$

Where,

- *similarity* is between 0.0 and 1.0. (0.0) means the strings do not have anything in common and (1.0) means the strings are exactly equal.

- (a) is the number of the first-string bigrams, and (b) is the number of the second-string bigrams.

A bigram means two adjacent characters from a string. For example, the set of bigrams of string “Moses” is {“Mo”, “os”, “se”, “es”} and a = 4. Likewise, the set of bigrams of string “Mosa” is {“Mo”, “os”, “sa”} and b = 3. Therefore,

$$similarity = \frac{2(a \ \cap b)}{(a+b)} = \frac{2(2)}{(7)} = 0.6$$

The following algorithm was used three times to measure the similarity between all concepts in the three Quranic datasets:

```
//Algorithm of Quranic concepts matching using Fuzzy string
//for measuring the similarity between Quranic concepts

Start
 input dataset1  //contains the list of concepts
 input dataset2  //contains the list of concepts
 output similarity_list

 for con1 in dataset1
   for con2 in dataset2
     list_bg1 = bigrams(con1)
     //list_bg1 is a list generated by bigrams function.
     list_bg2 = bigrams(con2)
     // list_bg2 is a list generated by bigrams function.
     similarity=2(size (list_bg1 intersect list_bg2)) divided by
                (size(list_bg1) + size(list_bg2))
     if (similarity > 0.4)
       similarity_list.add(con1, con2, similarity)
     end if
   end for
 end for
End
```

The algorithm returned the output lists containing similar pairs of concepts with similarity value between (0.40) and (1.0).  This is mean that, the threshold of similarity value  is at least two similar bigrams between two concepts.

After that, the lists were reviewed manually to verify the matching between similar concepts. Finally, the Arabic and English labels of each matched pair of concepts were corrected to have the same Arabic and English names. Table 10 presents a summary of the fuzzy string experiments.

**Table 10: Similar concepts between QT, QurAna and OQC using Simple Fuzzy string.**

| Dataset | # of Concepts | # Comparisons | # Matching time |
|---|---|---|---|
| QT and QurAna | 1150, 1050 | 1194669 | 339 |
| QT and OQC | 1150, 300 | 285606 | 100 |
| OQC and QurAna | 300, 1050 | 253135 | 93 |

After matching and merging similar concepts in these Qur'anic datasets. the new datasets are stored in a Database.

### 4.3.3 Developing ontology

In this stage, developing ontology includes the following consecutive steps: creating entities (classes, data properties, and object properties), testing ontology consistency, creating axioms, creating the Quran ontology documentation and deploying the Quran ontology.

classes are determined based on the 12 Quranic datasets. These classes were built depending on two main categories: the structure of the Quran, and the facts and knowledge mentioned in the Quran which are called concepts. The structure of the Quran describes information about the Quran chapters, verses, and words.

The total number of these classes are 78 classes which are created using Protégé 5.0. Figure 12 shows the hierarchy of the 78 Quranic classes in the English language, and figure 13 demonstrates the same 78 classes in the Arabic language.

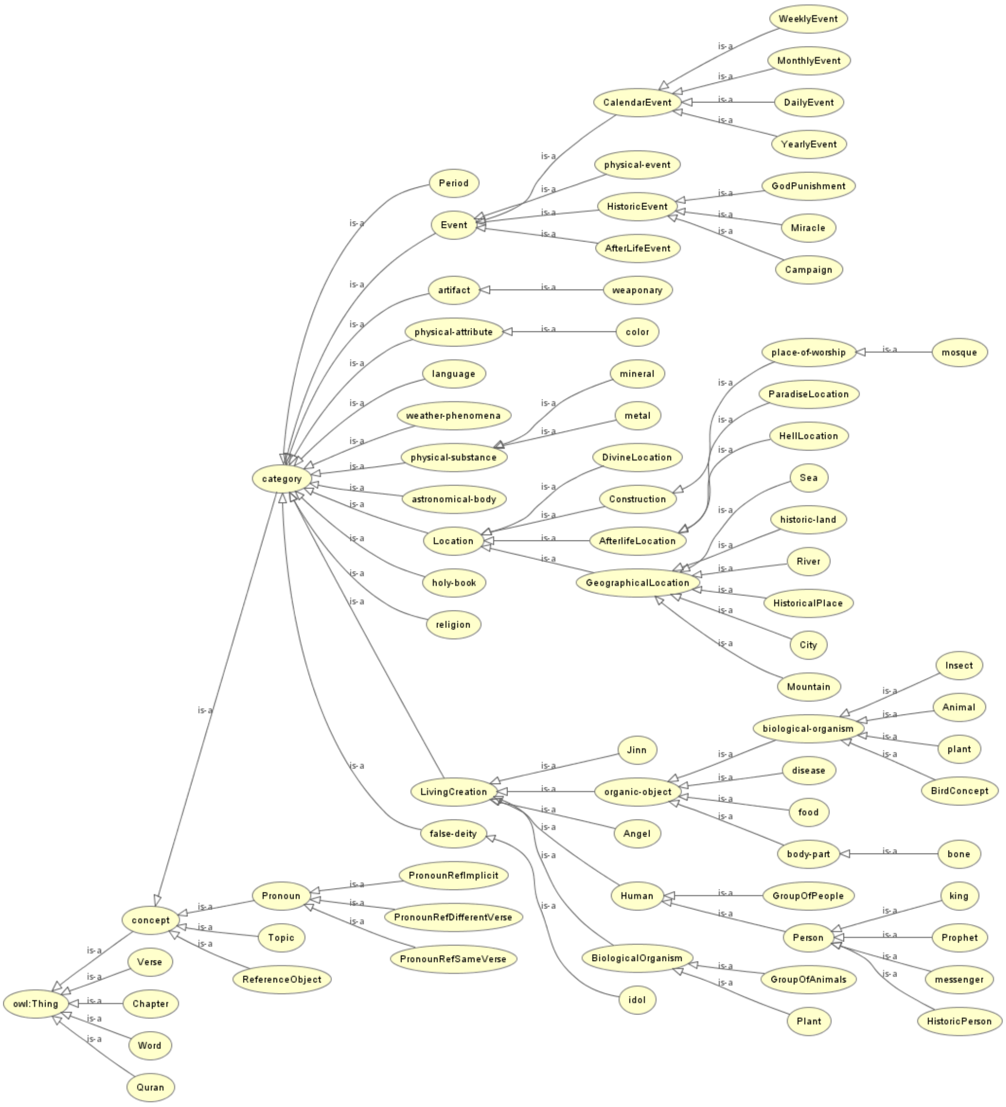


**Figure 12: Hierarchy of the English Quranic classes**

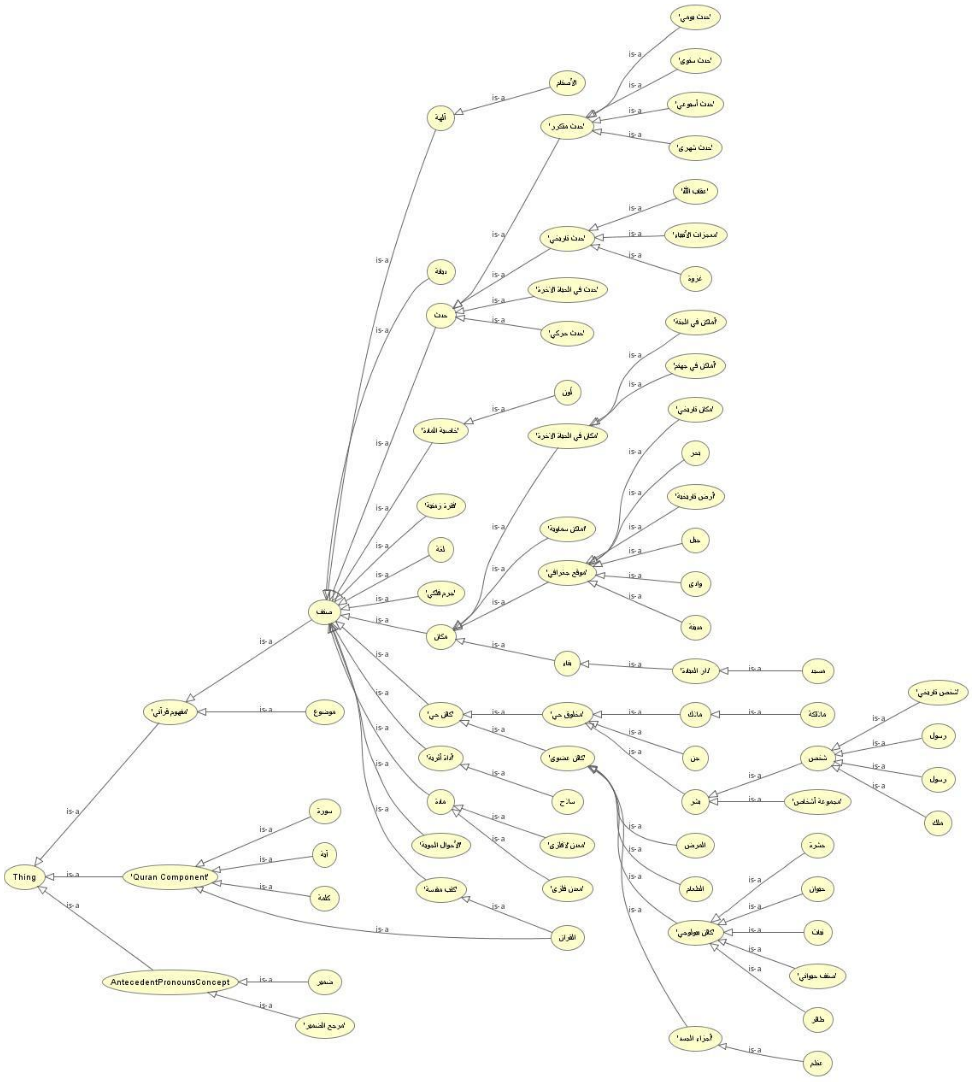


**Figure 13: Hierarchy of the Arabic Quranic classes**

Then, the data properties and object properties are created. The object property relates an object to an object (like verses to their chapter ), while the datatype property assigns a data value to an object(such as a 'verse index number' to a verse). Table 11 shows 'revealedAfter' as an example of object properties from the Quran ontology.

**Table 11: Declaration of object property ‘revealed after’**

| Object Property: revealedAfter | |
|---|---|
| rdfs:label "Revealed after" @en | rdfs:label "نزلت بعد" @ar |
| Domains: Chapter | Ranges: Chapter |
| Inverses: revealedBefore | Example:<br>Surah_Aal-i-Imraan<br>***revealedAfter***<br>Surah_Al-Anfaal |
| <!-- Resource : http://QuranSemanticData.com/Resource/<br>From Quran ontology owl:xml file --><br>  <owl:ObjectProperty rdf:about="&Resource;revealedAfter"><br>    <rdfs:label xml:lang="ar">نزلت بعد</rdfs:label><br>    <rdfs:label xml:lang="en">Revealed after</rdfs:label><br>    <rdfs:domain rdf:resource="&Resource;Chapter"/><br>    <rdfs:range rdf:resource="&Resource;Chapter"/><br>  </owl:ObjectProperty> | |

**Table 12: Declaration of Datatype property ‘word Meaning'**

| DatatypeProperty : wordMeaning | |
|---|---|
| rdfs:label "Word Meaning" @en | rdfs:label "معنى الكلمة" @ar |
| Domains: Word | rdfs:comment " the Arabic-Arabic meaning of a Quran word "(xsd:string) |
| <! -- http://QuranSemanticData.com/Resource/WordMeaning --><br>  <owl:DatatypeProperty rdf:about="&Resource;wordMeaning"><br>    <rdfs:label xml:lang="ar">معنى الكلمة</rdfs:label><br>    <rdfs:label xml:lang="en">Word Meaning</rdfs:label><br>    <rdfs:domain rdf:resource="&Resource;Word"/><br>    <rdfs:range rdf:resource="&rdfs;Literal"/><br>    <rdfs:subPropertyOf rdf:resource="&owl;topDataProperty"/><br>  </owl:DatatypeProperty> | |
| **Word**: ‘Their prayer’ **wordMeaning**: ‘دُعاؤهم’ | |

Table 12 shows the details of the Quran datatype property ‘word meaning’.

After that, this ontology was logically evaluated in which the consistency of the new Quran ontology is verified using FaCT++ and HermiT[72] reasoners provided by Protégé.

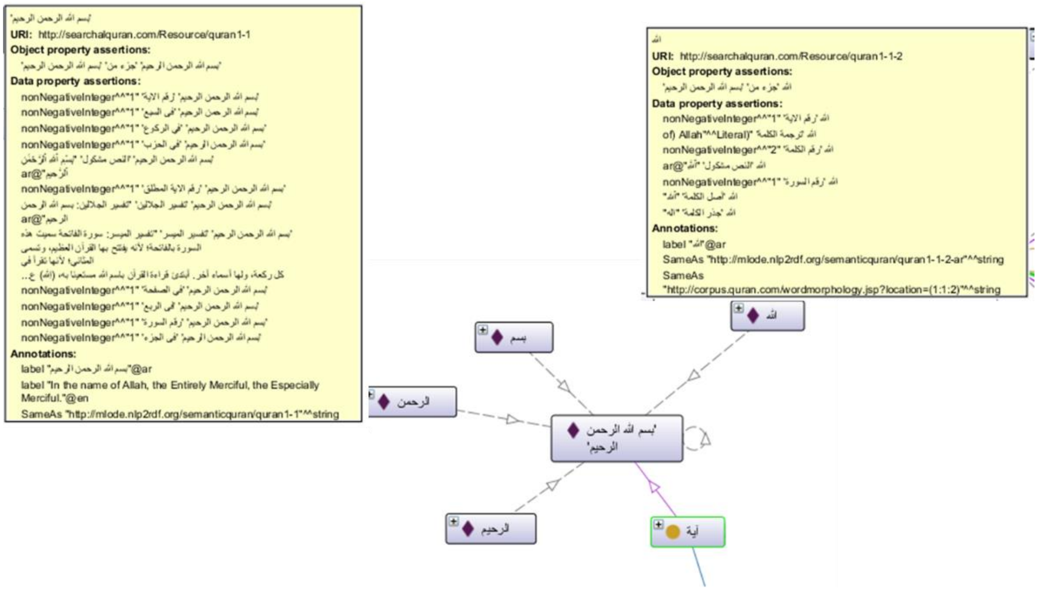
Then all instances (86588 individuals) of classes were fed from excel sheets to the Quran ontology by using ‘create axioms from excel book’ protégé tool. This step result in, a new Quran ontology which has more than 1,070,000 triples, 78 classes, 51 object properties, 34 data properties and 86588 individuals. Figure 14 shows the graph of the first verse in the Quran ‘بسم الله الرحمن الرحيم’and how it links to its words and parent class.

After that, the ontology documentation was created using OWLDOC protégé plugin.

Finally, all documentation of this ontology was deployed to linked open data cloud[73] and ‘*http://quransemanticdata.com/Resource/’*.

[72] *http://www.hermit-reasoner.com/*

[73] *https://lod-cloud.net/dataset/quransemanticdata*



'بسم الله الرحمن الرحيم'
URI: http://searchalquran.com/Resource/quran1-1
**Object property assertions:**
'بسم الله الرحمن الرحيم' 'جزء من' 'بسم الله الرحمن الرحيم'
**Data property assertions:**
'بسم الله الرحمن الرحيم' 'رقم الاية' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'فى السبع' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'فى الركوع' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'فى الحزب' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'النص مشكول' "بِسْمِ اللَّهِ الرَّحْمَنِ
الرَّحِيمِ"@ar
'بسم الله الرحمن الرحيم' 'رقم الاية المطلق' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'تفسير الجلالين' "تفسير الجلالين: بسم الله الرحمن
الرحيم"@ar
'بسم الله الرحمن الرحيم' 'تفسير الميسر' "تفسير الميسر: سورة الفاتحة سميت هذه
السورة بالفاتحة؛ لأنه يفتتح بها القرآن العظيم، وتسمى
المثاني؛ لأنها تقرأ في
كل ركعة، ولها أسماء أخر. أبتدئ قراءة القرآن بإسم الله مستعينا به، (اللهِ) ع..
'بسم الله الرحمن الرحيم' 'فى الصفحة' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'فى الربع' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'رقم السورة' "1"^^nonNegativeInteger
'بسم الله الرحمن الرحيم' 'فى الجزء' "1"^^nonNegativeInteger
**Annotations:**
label "بسم الله الرحمن الرحيم"@ar
label "In the name of Allah, the Entirely Merciful, the Especially Merciful."@en
SameAs "http://mlode.nlp2rdf.org/semanticquran/quran1-1"^^string

الله
URI: http://searchalquran.com/Resource/quran1-1-2
**Object property assertions:**
الله 'جزء من' 'بسم الله الرحمن الرحيم'
**Data property assertions:**
الله 'رقم الاية' "1"^^nonNegativeInteger
الله 'ترجمة الكلمة' "of) Allah"^^Literal)"
الله 'رقم الكلمة' "2"^^nonNegativeInteger
الله 'النص مشكول' "ٱللَّهِ"@ar
الله 'رقم السورة' "1"^^nonNegativeInteger
الله 'اصل الكلمة' "ٱللَّه"
الله 'جذر الكلمة' "اله"
**Annotations:**
label "الله"@ar
SameAs "http://mlode.nlp2rdf.org/semanticquran/quran1-1-2-ar"^^string
SameAs "http://corpus.quran.com/wordmorphology.jsp?location=(1:1:2)"^^string

**Figure 14 : Graph of the first verse in the Quran**

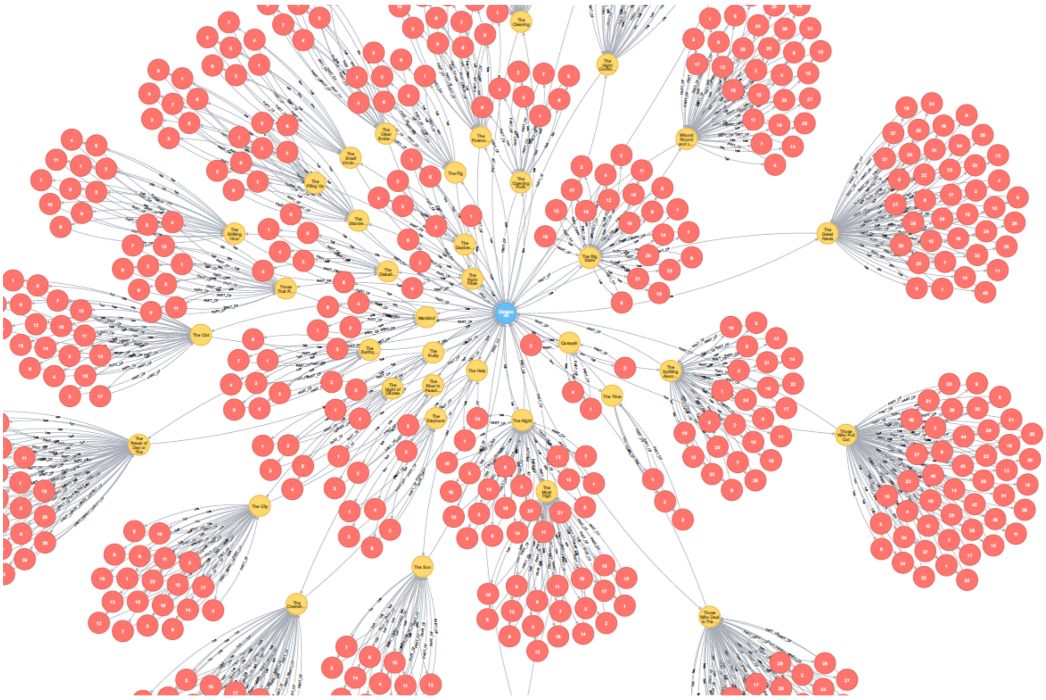


**Figure 15:The graph of Chapters and verses belonging to the Division (30) of the Quran**

## 4.4 Conclusion

This chapter reviewed the previous Quranic ontologies and then compared them against 14 criteria. Depending on this study, the current Arabic Quran ontologies have different: scopes, formats, entity names, and text language. Additionally, single Quranic ontology does not cover most of the knowledge in the Quran. Therefore, these ontologies needed to be increased, normalized, and combined with other Quran resources such as Quran words meanings.

Furthermore, some deficiencies in most of these ontologies were found, such as, validation and testing the consistency of these ontologies.

Therefore, a new Quranic ontology was developed to overcome the limitations of existing ontologies. The new Quran ontology was constructed from more than ten datasets related to the Quran domain. This ontology has more than 1,070,000 triples, 78 classes, 51 object properties, 34 data properties and 86588 individuals.

# Chapter 5
# Developing annotated Arabic Quranic question & answer corpus

This chapter will provide some background on corpus linguistics. Additionally, it will describe the different types of corpus linguistics and provides the background to Question and Answer (Q&A) corpora. The chapter emphasises the importance of Q&A corpora and summarises their use in Information Retrieval Systems (IR), such as for question classification and evaluation of the systems. It then details the construction of the Arabic Quranic Question and Answer Corpus (AQQAC) and describes the classification of the Arabic Quranic questions and answers using the Quran ontology. Finally, the chapter will introduce the semantic similarity model for the Quran Arabic words using Word Embedding techniques.

## 5.1 Introduction

The term corpus (plural corpora or corpuses) is defined as a collection of texts stored in electronic form that  depends on specific criteria of design, size and purpose (Kennedy, 1998).

### 5.1.1 Text corpora types

Text corpora are classified according to the source of their content, purpose and goal. Types of text corpora include: monolingual corpora, parallel corpora, multilingual corpora, comparable corpora, learner corpora, diachronic corpora and multimedia and specialised corpora (Hunston, 2002; Kennedy, 1998).

A monolingual corpus contains texts in only one language. In general, this corpus is tagged for parts of speech. The main uses of this corpus are:  checking the right usage of a word, finding the most natural word combinations and determining common patterns or a new trend in language.

A parallel corpus comprises a monolingual corpus and its translation, such as the Quran and its English translation. The two corpora need to be aligned at the level of paragraphs or sentences. A multilingual corpus is a parallel corpus which contains texts in several languages.

A comparable corpus is another example of a text corpus that comprises two or more corpora in different languages. The texts included in these corpora are related to the same topic but are not aligned and are not translations. The comparable corpus is used mainly by translators to study equivalency and differences between languages.

Another type of text corpus type is a learner corpus that contains a collection of texts produced by learners of a language. This corpus is used to identify learners' mistakes and problems when they are learning a foreign language. The Arabic Learner Corpus (ACL) is a collection of written and spoken material from learners of the Arabic Language. It comprises 282,732 words and  1585 resources produced by 942 individuals from 67 countries (Alfaifi, 2015).

An important text corpus type is the diachronic corpus which is a collection of texts from different periods of time. This historic corpus is used to trace the development of the language over time.

An additional form of text corpora is the multimedia corpus. This corpus consists of texts that are enriched with audio and video material, such as the British National Corpus that has the relevant recordings linked to the text of speeches.

Another type of corpus is the specialised corpus that comprises texts limited to one or more topics and domains. The main use of this corpus is to demonstrate how the aspect of a language is used. An example of such a corpus is the child language acquisition corpus.

## 5.2 Background of Question & Answer Corpora (QAC)

A question and answer corpus (QAC) is a collection of questions and answers that includes some metadata about each question and answer, such as the question asker, question topic, the date, question id, and question type. QACs are used to evaluate IR systems and classify questions.

The first known QAC is the Cranfield collection which was created in the late of 1960s by (Cleverdon, Mills, & Keen, 1966). This data set contained 225 queries and 1400 documents and was used to evaluate an IR system called the Cranfield II. This QAC has been used by researchers as testing data set since it was created.

A number of Q&A data sets have been built since, such as the CACM collection by Fox (1983).

At the start of the 1990s, the Text Retrieval Conference (TREC) was established to encourage researchers in the field of  information retrieval and to provide researchers with diverse large data collections to help evaluate their systems (Harmon, 1993). TREC-8 is an example of a Question and Answer corpus which contains 200 questions and answers. Most of these questions are mined from the FAQ-Finder systems logs. FAQ-FINDER is a natural language question-answer system using frequently-asked question files as its knowledge base.

Most QACs are collated from community forum websites, such as Yahoo Answers[74], wiki-Answers[75] and Quora[76]. These datasets are known as community Question-Answering (cQA) datasets.

An alternative kind of QAC is the Stanford Question Answering Dataset (SQuAD)[77]. This consists of 150,000 questions that were posed by crowd-workers on a collection of Wikipedia articles. The answer to each question forms part of the text of the corresponding article; some questions may remain un-answered (Rajpurkar, Zhang, Lopyrev, & Liang, 2016). Table 13 presents a list of most important Q&A corpora.

A closer look at the list of datasets given in Table 13 indicates that the datasets differ in structure, size and diversity of the questions. Some corpora consist of questions related to one specific topic; others contain questions on a range of different topics. Most of the datasets are not related to the Quran domain and are in English. To the best of the author’s knowledge, there are no Quranic Q&A corpora in Classical Arabic text available for public use.

[74] https://answers.yahoo.com/
[75] http://www.answers.com/
[76] https://www.quora.com/
[77] https://rajpurkar.github.io/SQuAD-explorer/

**Table 13: An overview of Question-Answering corpora**

| | **Resources in corpus** | **Use** |
|---|---|---|
| Berger, Caruana, Cohn, Freitag, and Mittal (2000) | Usenet FAQs[78] and call-centre dialogues[79] | Used to find correlations between questions and answers automatically |
| Bernhard and Gurevych (2009) | About 480,000 questions with their answers from Wiki-Answers[80] | Used to find answers |
| Yahoo! Web-scope L6-Dataset [81] | 4.5 M questions and their answers were collected from Yahoo! Answers on 10/2007. | Used validate answer extraction models |
| SQuAD2.0 (Rajpurkar, Jia, & Liang, 2018) | 150,000 questions posed by crowd-workers on Wikipedia articles | Used to test the performance of QA systems in finding answer from a certain passage and comparing them to answers written by individuals. |
| SemEval Task 3 (Nakov et al., 2015)[82] | 1500 pairs of questions and answers were collected from IslamWeb[83], and 2900 pairs were collected from the Qatar Living forum[84] | Used to categorise answers as bad, good, or possibly relevant to a question. Additionally, to answer a YES\NO question with yes, no, or unsure. |
| The Quora Dataset[85] | 400,000 possible question duplicate pairs. | Used to develop a model for identifying similar questions. |

[78] *ftp://rtfm.mit.edu.*

[79] *http://www.faqs.org.*

[80] *http://wiki.answers.com.*

[81] *https://webscope.sandbox.yahoo.com/catalog.php?datatype=l*

[82] *http://alt.qcri.org/semeval2015/task3/index.php?id=data-and-tools*

[83] *http://fatwa.islamweb.net/fatwa/index.php*

[84] *https://www.qatarliving.com/*

[85] *https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs*

## 5.3 Arabic Quranic Question & Answer corpus (AQQAC)

AQQAC is a collection of approximately 2224 questions and answers about the Quran. Each question and answer is annotated with the question ID, question word (particles), chapter number, verse number, question topic, question type, the Quran ontology concepts (Alqahtani & Atwell, 2018) and question source. Table 14 demonstrates an annotated question from AQQAC.

**Table 14: Example of annotated Arabic Quran question-answer**

| Tag | Text |
|---|---|
| **Question Number** | 685 |
| **Question** | من هو النبي الوزير حيث جعله الله تعالى في هذه المنزلة؟ وما الآية الدالة على ذلك؟<br>Who is the Prophet Allah assigned him as minister to another prophet ? which verse. |
| **Answer** | هارون عليه السلام قال تعالى: وَلَقَدْ آتَيْنا مُوسَى الْكِتابَ وَجَعَلْنا مَعَهُ أَخاهُ هارُونَ وَزِيراً أي معينا له وظهيرا.<br>Aaron, peace be upon him |
| **Question particles** | من (men): Who |
| **Location** | Chapter: Al Forgan, Verse: 35<br>[الفرقان: 35] |
| **Question topic** | النبي الوزير (the minister Prophet) |
| **Question class** | Rational: عاقل |
| **Question fine class** | Prophet: نبي |
| **The Quran ontology** | نبي , هارون : Prophet, Aaron |
| **Question source** | 1000QA |
| **Question type** | Fact |

The aim of this corpus is to provide a question & answer taxonomy for questions about the Quran. Additionally, this corpus might be used as a data set for testing and evaluating Islamic IR systems.

### 5.3.1 Methodology of developing AQQAC

The methodology for developing the AQQAC comprised four stages: the resources collection (will be described in section 5.3.2), the pre-processing of the corpus text (will be described in section 5.3.3), the extraction of questions and answers from the collected resources (will be described in section 5.3.4), and the annotation of the question and answer text (will be described in section 5.3.5). Figure 16 outline the methodology of developing the AQQAC step-by-step.

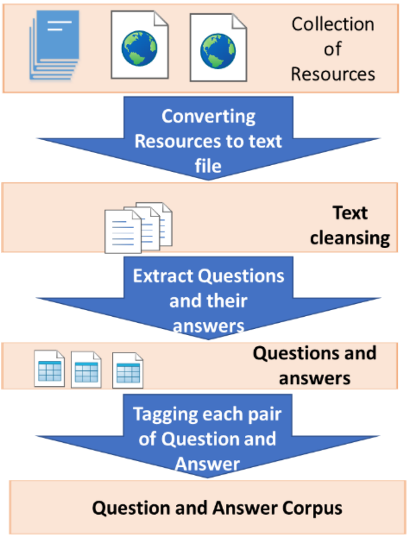

Collection of Resources
Converting Resources to text file
Text cleansing
Extract Questions and their answers
Questions and answers
Tagging each pair of Question and Answer
Question and Answer Corpus

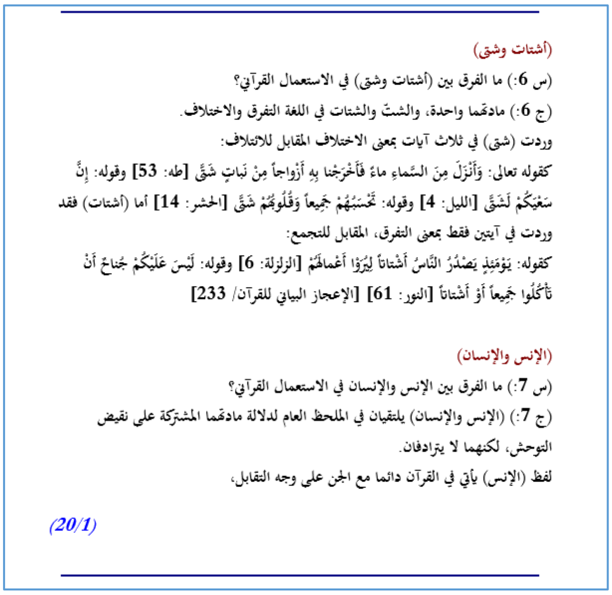
**Figure 16: AQQAC development stages**

### 5.3.2 Resources collection stage

The Quran question-answer datasets are collected from two sources: “1000 questions and answers on the Quran” and “Islam – the Quran and Tafseer website”. In this stage, the process of collecting each source is described in detail as follows.

### 5.3.2.1 1000 questions and answers on the Quran

This resource (***1000 Su'al Wa Jawab Fi ALKORAN***) was compiled by the famous Islamic scholar ‘Ashur (2001). This Arabic book contains 1000 questions and answers about the Quran, as in the example given in Figure 17. The electronic copy of this book was imported form Shamela,[86] which is a free library for electronic Islamic books. This eBook was converted into a text file with UTF-8 encoding format.



(أشتات وشتى)

(س 6:) ما الفرق بين (أشتات وشتى) في الاستعمال القرآني؟

(ج 6:) مادتهما واحدة، والشتّ والشتات في اللغة التفرق والاختلاف.

وردت (شتى) في ثلاث آيات بمعنى الاختلاف المقابل للائتلاف:

كقوله تعالى: وَأَنْزَلَ مِنَ السَّماءِ ماءً فَأَخْرَجْنا بِهِ أَزْواجاً مِنْ نَباتٍ شَتَّى [طه: 53] وقوله: إِنَّ سَعْيَكُمْ لَشَتَّى [الليل: 4] وقوله: تَحْسَبُهُمْ جَمِيعاً وَقُلُوبُهُمْ شَتَّى [الحشر: 14] أما (أشتات) فقد وردت في آيتين فقط بمعنى التفرق، المقابل للتجمع:

كقوله: يَوْمَئِذٍ يَصْدُرُ النَّاسُ أَشْتاتاً لِيُرَوْا أَعْمالَهُمْ [الزلزلة: 6] وقوله: لَيْسَ عَلَيْكُمْ جُناحٌ أَنْ تَأْكُلُوا جَمِيعاً أَوْ أَشْتاتاً [النور: 61] [الإعجاز البياني للقرآن/ 233]

(الإنس والإنسان)

(س 7:) ما الفرق بين الإنس والإنسان في الاستعمال القرآني؟

(ج 7:) (الإنس والإنسان) يلتقيان في الملحظ العام لدلالة مادتهما المشتركة على نقيض التوحش، لكنهما لا يترادفان.

لفظ (الإنس) يأتي في القرآن دائما مع الجن على وجه التقابل،

*(20/1)*

**Figure 17: two questions from 1000 Q&A about the Quran**

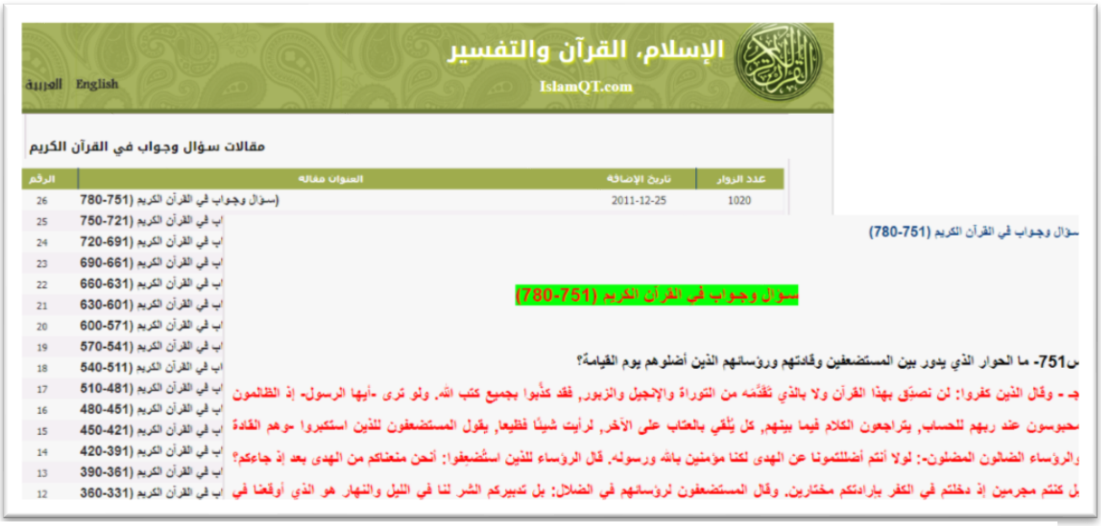
### 5.3.2.2 Islam – the Quran and Tafseer website "الإسلام القران والتفسير"

Islam – the Quran and Tafseer[87] is a website about the Quran that includes a description and a translation of the Quran and the reciting rules, the “Tajweed”. Additionally, this website has approximately 1200 questions and answers about the Quran in the Arabic language extracted from the Altabari Tafseer. These

[86] *http://shamela.ws/index.php/book/38133*

[87] *http://islamqt.com/*

questions were imported using the web scraping[88] technique. The scraping technique is used to extract data from a website. The scraping code is given in Appendix D.1. The imported texts were saved in one text file in UTF-8 format.



مقالات سؤال وجواب في القرآن الكريم

| الرقم | العنوان مقاله | تاريخ الإضافه | عدد الزوار |
|---|---|---|---|
| 26 | (سؤال وجواب في القرآن الكريم (780-751 | 2011-12-25 | 1020 |
| 25 | ب في القرآن الكريم (750-721 | | |
| 24 | ب في القرآن الكريم (720-691 | | |
| 23 | ب في القرآن الكريم (690-661 | | |
| 22 | ب في القرآن الكريم (660-631 | | |
| 21 | ب في القرآن الكريم (630-601 | | |
| 20 | ب في القرآن الكريم (600-571 | | |
| 19 | ب في القرآن الكريم (570-541 | | |
| 18 | ب في القرآن الكريم (540-511 | | |
| 17 | ب في القرآن الكريم (510-481 | | |
| 16 | ب في القرآن الكريم (480-451 | | |
| 15 | ب في القرآن الكريم (450-421 | | |
| 14 | ب في القرآن الكريم (420-391 | | |
| 13 | ب في القرآن الكريم (390-361 | | |
| 12 | ب في القرآن الكريم (360-331 | | |

ـؤال وجواب في القرآن الكريم (780-751)

سؤال وجواب في القرآن الكريم (780-751)

س751- ما الحوار الذي يدور بين المستضعفين وقادتهم ورؤسائهم الذين أضلوهم يوم القيامة؟

جـ - وقال الذين كفروا: لن نصدِّق بهذا القرآن ولا بالذي تَقَدَّمَه من التوراة والإنجيل والزبور, فقد كذَّبوا بجميع كتب الله. ولو ترى -أيها الرسول- إذ الظالمون
محبوسون عند ربهم للحساب, يتراجعون الكلام فيما بينهم, كل يُلْقي بالعتاب على الآخر, لرأيت شيئًا فظيعا, يقول المستضعفون للذين استكبروا -وهم القادة
والرؤساء الضالون المضلون-: لولا أنتم أضللتمونا عن الهدى لكنا مؤمنين بالله ورسوله. قال الرؤساء للذين استُضعِفوا: أنحن منعناكم من الهدى بعد إذ جاءكم؟
بل كنتم مجرمين إذ دخلتم في الكفر بإرادتكم مختارين. وقال المستضعفون لرؤسائهم في الضلال: بل تدبيركم الشر لنا في الليل والنهار هو الذي أوقعنا في

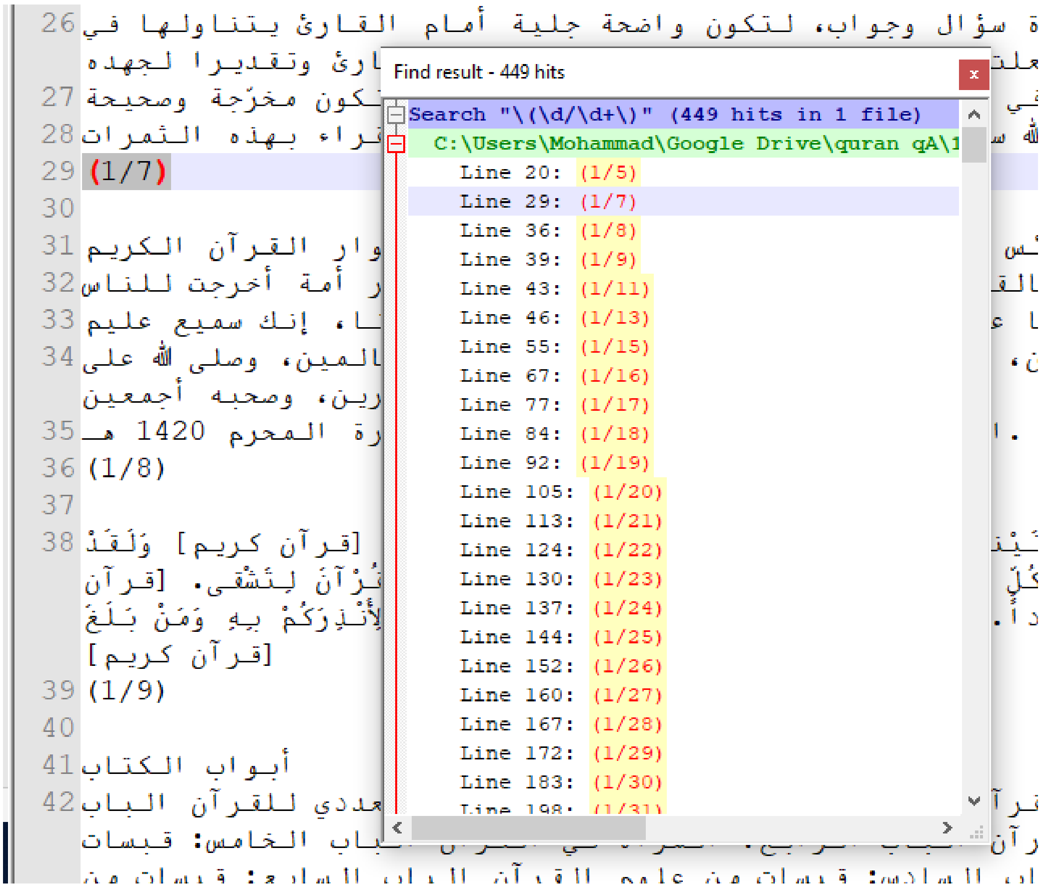
**Figure 18: A Question from IslamQT.com**

### 5.3.3 Pre-processing corpus text stage (text cleansing)

Data cleansing is the process of detection and removing errors and inconsistencies from data to improve the quality of data. Data cleaning can positively affect data analysis.

The pre-processing stage deals with elements that are not related to the text of the questions or answers including page numbers, headings, footnotes, non-Arabic letters and non-letter characters. The main challenge to remove these elements is that the different patterns of these elements need lots of  effort and time. Therefore, many regular expressions patterns were applied to the text to find and then, remove the unwanted elements. For example, in figure 19 the regular expression “\(\d/\d+\) ” was used to delete 449 page-numbers.

[88] https://en.wikipedia.org/wiki/Web_scraping



ة سؤال وجواب، لـتكون واضحة جلية أمام القارئ يتناولها في
ارئ وتقديرا لجهده
ـكون مخرّجة وصحيحة
قراء بهذه الثمرات
(1/7)

وار القرآن الكريم
ر أمة أخرجت للناس
ـاء، إنك سميع عليم
المين، وصلى الله على
رين، وصحبه أجمعين
رة المحرم 1420 هـ
(1/8)

[قرآن كريم] وَلَقَدْ
ُرْآنَ لِتَشْقَى. [قرآن
لِأُنْذِرَكُمْ بِهِ وَمَنْ بَلَغَ
[قرآن كريم]
(1/9)

أبواب الكتاب
عددي للقرآن الباب
باب الخامس: قبسات

Find result - 449 hits

```
Search "\(\d/\d+\)" (449 hits in 1 file)
  C:\Users\Mohammad\Google Drive\quran qA\1
    Line 20: (1/5)
    Line 29: (1/7)
    Line 36: (1/8)
    Line 39: (1/9)
    Line 43: (1/11)
    Line 46: (1/13)
    Line 55: (1/15)
    Line 67: (1/16)
    Line 77: (1/17)
    Line 84: (1/18)
    Line 92: (1/19)
    Line 105: (1/20)
    Line 113: (1/21)
    Line 124: (1/22)
    Line 130: (1/23)
    Line 137: (1/24)
    Line 144: (1/25)
    Line 152: (1/26)
    Line 160: (1/27)
    Line 167: (1/28)
    Line 172: (1/29)
    Line 183: (1/30)
```

**Figure 19: Search for page numbers in *1000 Question answers in the Quran* using regular expression tool**

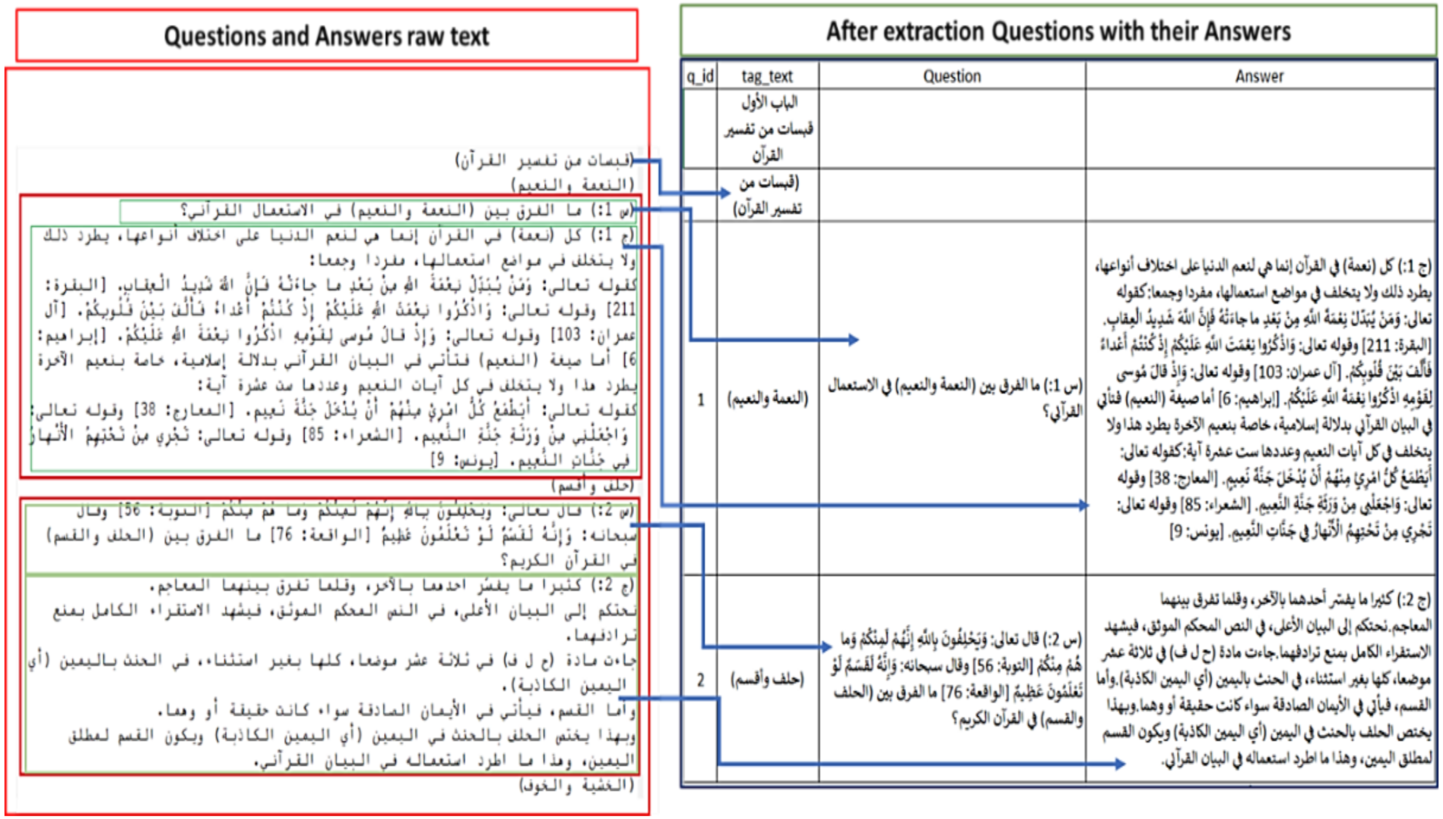
### 5.3.4 Extracting questions and answers stage

This stage involves extracting each question and answer from the unstructured Q&A documents and inputting them into spreadsheet of four columns. The columns are: the chapter title, question text, answer text, and question title. Figure 20 describes the spreadsheet file after processing the text file using the algorithm as follows:

```
// Extracting Questions and answers from raw text
// Input Question-Answer.txt
// Output output_file
Start
 Open output_file

 read input_file -> Question-Answer.txt
 for each line in input_file
    if(line starts with('الباب') ):
       // adding column1 in output file
```

```
          write in output_file('\n' + line)
     else if (line starts with('(س')):
      //adding column2 question text
      //where ‘;’ separates each column in   the same raw
          write in output_file (';' + line)
     else if (line starts with('(ج')):
          //adding column3 answer text
          write in output_file(';' + line)
     else if (regular expresion ('(\((.+)\))' match line)):
     // adding column4 ‘question title’
          write in output_file(';' + line)
     else:
          write in output_file( line)
 end for
 close output_file
End
```

This algorithm was written in python  programming language as exposed in Appendix D.2.



Questions and Answers raw text

(قبسات من تفسير القرآن)
(النعمة والنعيم)
(س 1:) ما الفرق بين (النعمة والنعيم) في الاستعمال القرآني؟
(ج 1:) كل (نعمة) في القرآن إنما هي لنعم الدنيا على اختلاف أنواعها، يطرد ذلك
ولا يتخلف في مواضع استعمالها، مفردا وجمعا:
كقوله تعالى: وَمَنْ يُبَدِّلْ نِعْمَةَ اللهِ مِنْ بَعْدِ ما جاءَتْهُ فَإِنَّ اللهَ شَدِيدُ الْعِقابِ. [البقرة:
211] وقوله تعالى: وَاذْكُرُوا نِعْمَتَ اللهِ عَلَيْكُمْ إِذْ كُنْتُمْ أَعْداءً فَأَلَّفَ بَيْنَ قُلُوبِكُمْ. [آل
عمران: 103] وقوله تعالى: وَإِذْ قالَ مُوسى لِقَوْمِهِ اذْكُرُوا نِعْمَةَ اللهِ عَلَيْكُمْ. [إبراهيم:
6] أما صيغة (النعيم) فتأتي في البيان القرآني بدلالة إسلامية، خاصة بنعيم الآخرة
يطرد هذا ولا يتخلف في كل آيات النعيم وعددها ست عشرة آية:
كقوله تعالى: أَيَطْمَعُ كُلُّ امْرِئٍ مِنْهُمْ أَنْ يُدْخَلَ جَنَّةَ نَعِيمٍ. [المعارج: 38] وقوله تعالى:
وَاجْعَلْنِي مِنْ وَرَثَةِ جَنَّةِ النَّعِيمِ. [الشعراء: 85] وقوله تعالى: تَجْرِي مِنْ تَحْتِهِمُ الْأَنْهارُ
فِي جَنّاتِ النَّعِيمِ. [يونس: 9]
(حلف وأقسم)
(س 2:) قال تعالى: وَيَحْلِفُونَ بِاللهِ إِنَّهُمْ لَمِنْكُمْ وَما هُمْ مِنْكُمْ [التوبة: 56] وقال
سبحانه: وَإِنَّهُ لَقَسَمٌ لَوْ تَعْلَمُونَ عَظِيمٌ [الواقعة: 76] ما الفرق بين (الحلف والقسم)
في القرآن الكريم؟
(ج 2:) كثيرا ما يفسّر احدهما بالآخر، وقلما تفرق بينهما المعاجم.
نحتكم إلى البيان الأعلى، في النص المحكم الموثق، فيشهد الاستقراء الكامل بمنع
ترادفهما.
جاءت مادة (ح ل ف) في ثلاثة عشر موضعا، كلها بغير استثناء، في الحنث باليمين (أي
اليمين الكاذبة).
وأما القسم، فيأتي في الأيمان الصادقة سواء كانت حقيقة أو وهما.
وبهذا يختص الحلف بالحنث في اليمين (أي اليمين الكاذبة) ويكون القسم لمطلق
اليمين، وهذا ما اطرد استعماله في البيان القرآني.
(الخشية والخوف)

After extraction Questions with their Answers

| q_id | tag_text | Question | Answer |
|---|---|---|---|
| | الباب الأول قبسات من تفسير القرآن | | |
| | (قبسات من تفسير القرآن) | | |
| 1 | (النعمة والنعيم) | (س 1:) ما الفرق بين (النعمة والنعيم) في الاستعمال القرآني؟ | (ج 1:) كل (نعمة) في القرآن إنما هي لنعم الدنيا على اختلاف أنواعها، يطرد ذلك ولا يتخلف في مواضع استعمالها، مفردا وجمعا:كقوله تعالى: وَمَنْ يُبَدِّلْ نِعْمَةَ اللهِ مِنْ بَعْدِ ما جاءَتْهُ فَإِنَّ اللهَ شَدِيدُ الْعِقابِ. [البقرة: 211] وقوله تعالى: وَاذْكُرُوا نِعْمَتَ اللهِ عَلَيْكُمْ إِذْ كُنْتُمْ أَعْداءً فَأَلَّفَ بَيْنَ قُلُوبِكُمْ. [آل عمران: 103] وقوله تعالى: وَإِذْ قالَ مُوسى لِقَوْمِهِ اذْكُرُوا نِعْمَةَ اللهِ عَلَيْكُمْ. [إبراهيم: 6] أما صيغة (النعيم) فتأتي في البيان القرآني بدلالة إسلامية، خاصة بنعيم الآخرة يطرد هذا ولا يتخلف في كل آيات النعيم وعددها ست عشرة آية:كقوله تعالى: أَيَطْمَعُ كُلُّ امْرِئٍ مِنْهُمْ أَنْ يُدْخَلَ جَنَّةَ نَعِيمٍ. [المعارج: 38] وقوله تعالى: وَاجْعَلْنِي مِنْ وَرَثَةِ جَنَّةِ النَّعِيمِ. [الشعراء: 85] وقوله تعالى: تَجْرِي مِنْ تَحْتِهِمُ الْأَنْهارُ فِي جَنّاتِ النَّعِيمِ. [يونس: 9] |
| 2 | (حلف وأقسم) | (س 2:) قال تعالى: وَيَحْلِفُونَ بِاللهِ إِنَّهُمْ لَمِنْكُمْ وَما هُمْ مِنْكُمْ [التوبة: 56] وقال سبحانه: وَإِنَّهُ لَقَسَمٌ لَوْ تَعْلَمُونَ عَظِيمٌ [الواقعة: 76] ما الفرق بين (الحلف والقسم) في القرآن الكريم؟ | (ج 2:) كثيرا ما يفسّر أحدهما بالآخر، وقلما تفرق بينهما المعاجم.نحتكم إلى البيان الأعلى، في النص المحكم الموثق، فيشهد الاستقراء الكامل بمنع ترادفهما.جاءت مادة (ح ل ف) في ثلاثة عشر موضعا، كلها بغير استثناء، في الحنث باليمين (أي اليمين الكاذبة).وأما القسم، فيأتي في الأيمان الصادقة سواء كانت حقيقة أو وهما.وبهذا يختص الحلف بالحنث في اليمين (أي اليمين الكاذبة) ويكون القسم لمطلق اليمين، وهذا ما اطرد استعماله في البيان القرآني. |

Figure 20: Extraction of questions and answers from raw text

### 5.3.5 Annotating the Arabic Quranic Question-Answer text stage

After extracting question-answers pairs in a spreadsheet file, each pair of question and answer will be tagged by: question number, answer location in the Quran (Quran chapter, verse), question word, question topic, the Quran ontology concept, and the source of the question. Table 15 describes the meaning of each tag in the AQQAC tag-set.

**Table 15: AQQAC tag set**

| Tag | Tag Manning |
|---|---|
| **Question Number** | The question number |
| **Question** | The text of the query |
| **Answer** | The text of corresponding answers to the question |
| **Question particles** | Arabic question particles which appear in a question |
| **Location** | Location of the answer in the Quran. This location can be assigned according to the name of the Quran chapter and its verse such as [chapter: verse] |
| **Question topic** | The topic of a question |
| **Question Class** | Quranic hierarchy class a question belongs to |
| **Question Fine class** | Quranic hierarchy sub class the question belongs to |
| **The Quran ontology concept** | The Quran ontology concept the question belongs to |
| **Question source** | The original source of the question |
| **Question type** | The type of answer to the question; i.e. Factual or Descriptive |

The tagging process of AQQAC has three steps: “annotating a Question-Answer text with its location in the Quran”, “annotating each question-answer text with the Quran ontology concepts” and, “annotating Quranic Question & Answers with question class”.

#### 5.3.5.1 Annotating a Question-Answer text with its location in the Quran

Each answer of a question consists of a Quranic verse and additional description text. In this stage, a verse number is extracted from the text of each answer. After that, the extracted verse number is added as a location tag. The verse numbers were extracted using an algorithm containing regular expression techniques. The code of the algorithm is shown in Appendix D.3.

#### 5.3.5.2 Annotating each question-answer text with the Quran ontology concepts

The annotating stage of the Quranic questions went through two steps. the first step is that, tagging each questions with Quranic ontology concepts found in the same verse in both question-answer and the Quranic ontology. The second step is selecting the most relevant concept from group of concepts linked to each question.

In the first step, each question is annotated with the Quran ontology concepts (previously described in section 4.3) that are linked to the Quran verses appearing in the answer of this question. In this Quranic ontology, each Quran verse is linked to at least one Quranic ontology concept. Therefore, a question-answer pair is tagged with the Quran ontology concepts which are linked to the verses appearing in the answer. However, a verse could be associated with more than one Quran concept, which is not necessarily related to the question that has the same verse as its answer. For example, Table 16 presents both the associated and the non-associated Quranic concepts in relation to a particular question.

**Table 16: Question annotated by the Quran ontology concepts**

| | |
|---|---|
| Question | ما هي معجزات عيسى عليه السلام؟ اذكر الآية<br>What are the miracles of Jesus, peace be upon him? Mention the verse? |
| Answer from the Quran<br>(verse 5:110) | إِذْ قَالَ اللَّهُ يَا عِيسَـــى ابْنَ مَرْيَمَ اذْكُرْ نِعْمَتِي عَلَيْكَ وَعَلَىٰ وَالِدَتِكَ إِذْ أَيَّدتُّكَ بِرُوحِ الْقُدُسِ تُكَلِّمُ النَّاسَ فِي الْمَهْدِ وَكَهْلًا ۖوَإِذْ عَلَّمْتُكَ الْكِتَابَ وَالْحِكْمَةَ وَالتَّوْرَاةَ وَالْإِنجِيلَ ۖوَإِذْ تَخْلُقُ مِنَ الطِّينِ كَهَيْئَةِ الطَّيْرِ بِإِذْنِي فَتَنفُخُ فِيهَا فَتَكُونُ طَيْرًا بِإِذْنِي ۖوَتُبْرِئُ الْأَكْمَهَ وَالْأَبْرَصَ بِإِذْنِي ۖوَإِذْ تُخْرِجُ الْمَوْتَىٰ بِإِذْنِي ۖوَإِذْ كَفَفْتُ بَنِي إِسْـــرَائِيلَ عَنكَ إِذْ جِئْتَهُم بِالْبَيِّنَاتِ فَقَالَ الَّذِينَ كَفَرُوا مِنْهُمْ إِنْ هَٰذَا إِلَّا سِحْرٌ مُّبِينٌ - 5:110 |
| English translation<br>of the answer | [The Day] when Allah will say, "O Jesus, Son of Mary, remember My favour upon you and upon your mother when I supported you with the Pure Spirit and you spoke to the people in the cradle and in maturity; and [remember] when I taught you writing and wisdom and the Torah and the Gospel; and when you were designed from clay [that was] in the form of a bird with My permission, then you breathed into it, and it became a bird with My permission; and you healed the blind and the leper with My permission; and when you brought forth the dead with My permission; and when I restrained the Children of Israel from [killing] you when you came to them with clear proof and those who disbelieved among them said, "This is not but obvious magic."[89]<br>[Chapter 5: verse 110] |
| All Concepts of the Quran ontology related to the verse | التوراة - عيسى -الإنجيل – الأبرص - هيئة الطير - الله - بني اسرائيل – مريم – الطين – إسرائيل - الطير - الكفار<br>Bird, Clay, Gospel, Israel, Jesus, Leprosy, Maryam, Torah, Allah, form of a bird, the infidels, the Children of Israel |
| Concepts related to the question | التوراة - عيسى -الإنجيل – الأبرص - هيئة الطير - الطير<br>Bird, Gospel, Jesus, Leprosy, Torah, form of a bird |
| Concepts not related to the question | الله - بني اسرائيل – مريم – الطين – إسرائيل - الكفار<br>Clay, Israel, Maryam, Allah, the infidels, the Children of Israel |

After the questions are tagged with the corresponding Quran ontology concepts, selecting the most relevant Quran concepts to each question has two steps: using cosine similarity and manual selecting.

The first step is using cosine similarity to select the most relevant Quranic concept from set of concepts tagging a question. The selection is based on the highest value of cosine similarity between the question's nouns and these Quranic concepts. The cosine similarity is found by calculating the similarity between two non-zero vectors of an inner product space which computes the cosine of the angle between these vectors (Manning et al., 2008). The cosine of these non-zero vectors can be obtained by using the Euclidean dot product formula[90]:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

where, the vectors A and B are usually term frequency vectors of a certain text. The Cosine similarity code applied to extract this information is presented in Appendix D.4

Table 17 shows the highest score of cosine similarity between a question and a selected Quran concept. The total number of tagged questions that are the best match with the Quran concepts using cosine similarity is 882 out of 2224. The percentage of the annotated questions is around 40 percent of the questions in AQQAC.

The second step is manual tagging of the questions with Quranic concepts. 1696 questions were annotated with the Quran ontology concepts, which means that, in total, 79 percent of the questions in AQQAC were annotated.

[89] *https://quran.com/5/110?translations=20*

[90] *https://en.wikipedia.org/wiki/Cosine_similarity*

**Table 17: the highest score of cosine similarity between a question and the selected Quran concept.**

| Question ID | concepts | Similarity (0-1) | intersection between stemmed word Q&A and Ontology | intersection between Q&A and Ontology |
|---|---|---|---|---|
| 345 | الأشهر الحرم | 0.7135 | حرم ,شهر | الحرم ,الأشهر |
| 32 | القتال في الأشهر الحرم | 0.666 | قتل ,في ,شهر ,حرم | الحرم ,في ,القتال ,الأشهر |
| 799 | الصدق | 0.663 | صدق | صدق |
| 338 | براءة الله ورسوله من المشركين | 0.652 | برء ,مشركين ,الل ,رسل ,من | ورسوله ,الله ,من ,براءة |
| 1111 | نفخهم في الصور | 0.632 | في ,نفخ ,صور | الصور ,في |
| 1150 | الطامة الكبرى | 0.572 | كبرى ,طمة | الطامة ,الكبرى |
| 1077 | حملهم العرش | 0.555 | عرش ,حمل | عرش ,حمل |
| 450 | عمل الكفار لا ينفعهم يوم القيامة | 0.516 | عمل ,كفر ,لا ,يوم ,نفع ,قيامة | الكفار ,القيامة ,ينفعهم ,يوم ,لا |
| 582 | من يعبد الله على حرف | 0.508 | حرف ,الل ,من ,على ,عبد | يعبد ,الله ,من ,على |
| 679 | نفخهم في الصور | 0.479 | في ,نفخ ,صور | في |
| 485 | جنات عدن | 0.453 | عدن ,جنت | عدن ,جنات |

### 5.3.5.3 Annotating AQQAC with Quranic question classes

the question classes in table 19 were used to classify 2224 questions from the AQQAC (described further in section 5.3.6). Each question was manually tagged by the appropriate question class. Table 18 shows the distribution of the 2224 questions over the Quran question hierarchy. Coarse classes are in bold and are followed by their fine classes. The Count column shows the number of questions in each class.

**Table 18: The distribution of 2224 questions over the Quran question hierarchy.**

| CLASS | fine class | | count |
|---|---|---|---|
| Allah | name | الله | 1 |
| Number | | | 94 |
| | occurrence | تكرار | 53 |
| | count | عدد | 24 |
| | period | فترة | 17 |
| ENTITY | | | 774 |
| | test | ابتلاء | 1 |
| | other | آخر | 3 |
| | false-deity | آلهة | 2 |
| | verse | آية | 545 |
| | body | الجسم | 1 |
| | astronomical | الكون أو جزء منه | 3 |
| | craft | حرفة | 1 |
| | animal | حيوان | 1 |
| | religion | دين | 1 |
| | chapter | سورة | 90 |
| | strong desire | شهوة | 1 |
| | food | طعام | 4 |
| | weather | ظاهرة الطقس | 2 |
| | punishment | عقاب | 20 |
| | sign | علامة | 10 |
| | battle | غزوة | 6 |
| | story | قصة | 40 |
| | artifact | قطعة أثرية | 1 |
| | holy-book | كتاب مقدس | 5 |
| | language | لغة | 2 |
| | substance | مادة | 3 |
| | proverb | مثل | 2 |
| | term | مصطلح | 9 |
| | miracle | معجزة | 7 |
| | rewards | مكافأة | 2 |
| | advice | موعظة | 8 |
| | plant | نبات | 3 |
| Event | | | 22 |
| | calendar | التقويم | 2 |
| | afterlife | حدث الاخرة | 6 |
| | historic | حدث تاريخي | 7 |
| | physical | حدث حركي | 7 |
| Description | | | 983 |
| | other | آخر | 7 |
| | definition | تعريف | 31 |
| | law | حكم | 9 |
| | moral | خلق | 6 |
| | proof | دليل | 8 |
| | reason | سبب | 126 |
| | description | شرح | 371 |
| | condition | شرط | 8 |
| | attribute | صفة | 18 |
| | request | طلب | 5 |
| | difference | فرق | 28 |
| | favour | فضل | 5 |
| | list | قائمة | 7 |
| | speech | قول | 37 |
| | meaning | معنى | 279 |
| | manner | منهج | 31 |
| | result | نتيجة | 7 |
| Rational | | | 224 |
| | person | إنسان | 42 |
| | ghost | جني | 1 |
| | angels | ملائكة | 3 |
| | angel | ملك | 4 |
| | people | ناس | 84 |
| | prophet | نبي | 73 |
| | description | وصف | 17 |
| Location | | | 20 |
| | other | آخر | 1 |
| | house | بيت | 1 |
| | mountain | جبل | 2 |
| | worship place | دار العبادة | 2 |
| | city | قرية | 8 |
| | afterlife | موقع في الاخرة | 5 |
| | other | اخرى | 1 |
| YES/NO | | | 102 |
| | choice | اختياري | 1 |
| | yes-no | نعم-لا | 101 |
| | | | |

### 5.3.6 Quranic Question & Answers classification

The Quran question taxonomy is based on the combination of the Quran ontology concepts developed by Alqahtani and Atwell (2018) and question classification classes developed by Li and Dan (2002). This classification is concerned with the mapping of a question about the Quran into various semantic categories to predict a correct answer (previously described in section 2.2.4). The Quran question taxonomy uses two layered question hierarchy which contains eight coarse classes (entity, God, rational, description, location, number, event, polar question) and 85 fine classes. Table 19 explains in details the coarse classes and their fine classes of this classification system.

**Table 19: The Quran question taxonomy**

| **Class** | **Fine Class** | **Definition** | **صنف** | **تعريف** |
|---|---|---|---|---|
| ***entity*** | - | ***concrete or abstract object*** | ***كيان*** | |
| ***entity*** | **astronomy** | Astronomical body such as sun, moon. | **الكون أو جزء منه** | الكون ومكوناته |
| ***entity*** | **artifact** | An object made by a human being, typically an item of cultural or historical interest such as tools, production, instruments. | **قطعة أثرية** | اي آلة او اداة مثل سفينة نوح |
| ***entity*** | **holy-book** | Sacred texts of different religions such as the Quran and the Torah. | **كتاب مقدس** | الكتب السماوية |
| ***entity*** | **false-deity** | A false idol, deity. | **آلهة** | كل معبود من دون أو مع الله: آلة، صنم |
| ***entity*** | **insect** | A small arthropod animal that has six legs and wings such as bug, roach or worm. | **الحشرة** | الحشرات |
| ***entity*** | **animal** | A living creature that feeds on organic substance, typically having particular sense organs and nervous system except human and insect. Example, wolf and mouse. | **حيوان** | الحيوانات |
| ***entity*** | **body** | Organs of the human body. | **الجسم** | الجسم وأجزاءه |
| ***entity*** | **colour** | Colours such as red and green. | **اللون** | الألوان |
| ***entity*** | **currency** | A system of money in general use in ancient times such as Dinar. | **عملة** | النقود كالدرهم والدينار |

| Class | Fine Class | Definition | صنف | تعريف |
|---|---|---|---|---|
| ***entity*** | **medicine** | Diseases and medicine. | **الامراض والعلاج** | الأمراض والادوية لها |
| ***entity*** | **food** | Any nutritious substance that people or animals eat or drink. | **طعام** | الطعام من فواكه وخضار وحبوب |
| ***entity*** | **language** | A system of communication used by a particular community or country such as the Arabic and English language. | **اللغة** | لغة تحدث مثل اللغة العربية |
| ***entity*** | **letter** | Letters that are mentioned in the opening of some Quranic chapters such as (حم , ق, ص). | **حرف** | الحروف المتقطعة في بدايات سور القران |
| ***entity*** | **other** | Other entities not declared in the fine class list. | **آخر** | أي كيان آخر |
| ***entity*** | **plant** | A living organism of the kind exemplified by trees, shrubs, herbs, grasses, ferns, and mosses, typically growing in a permanent site, absorbing water and inorganic substances through its roots, such as flowers and herbs. | **نبات** | اي نبات من شجر ونحوه |
| ***entity*** | **religion** | The belief in and worship of a superhuman controlling power, such as Islam. | **دين** | الاديان السماوية |
| ***entity*** | **Sport** | An activity involving physical exertion and skill such as gambling. | **رياضة** | الرياضة والالعاب مثل الفروسية والقمار |
| ***entity*** | **substance** | Materials, elements and substances. | **مادة** | المواد مثل المعادن وغيرها |
| ***entity*** | **sign** | Sign indicates the probable presence or occurrence of something else such as the crack of doom. | **علامة** | علامة تدل على شيء أو وقوعه |
| ***entity*** | **term** | A word or phrase used to describe a thing or to express a concept. | **مصطلح** | مصطلح او اسم لشيء ما |
| ***entity*** | **chapter** | The Quran chapter | **سورة** | سور القران |
| ***entity*** | **verse** | The Quran verse | **آية** | آية من آيات القران |
| ***entity*** | **punishment** | The infliction or imposition of a penalty as retribution for an offense. | **عقاب** | عقاب أو عذاب |
| ***entity*** | **story** | Historical story that is mentioned in the Quran | **قصة** | قصص القران |

| Class | Fine Class | Definition | صنف | تعريف |
|---|---|---|---|---|
| ***entity*** | **strong desire** | Strong desires | **شهوة** | **شهوة ـ رغبة** |
| ***entity*** | **proverb** | A short, well-known pithy saying, stating a general truth or piece of advice. | **مثل** | مَثل ـ حكمة |
| ***entity*** | **battle** | Battle, war. | **غزوة** | معركة ـ غزوة ـ سرية |
| ***entity*** | **advice** | Instruction- advice- sermon. | **موعظة** | نصيحة ـ وصية ـ موعظة ـ تذكرة |
| ***entity*** | **sin** | An immoral act that is a transgression against divine law. | **إثم** | خَطِيئَة؛ ذنب؛ سَيِّئَة |
| ***entity*** | **rewards** | Rewards or gift. | **جائزة او مكافأة** | جائزة ـ مكافأة ـ هبة |
| ***entity*** | **test** | Trial; a tremendous trial by Allah in which the faith of a true believer is being tested. | **ابتلاء** | ابتلاء ـ اختبار ـ امتحان |
| ***entity*** | **weather-phenomena** | Rain, wind, thunder, lightning , cloud, dust | **ظاهرة الطقس** | احوال وظواهر الطقس: الرعد والبرق |
| ***entity*** | **miracle** |  | **معجزة** | معجزة او آية لنبي |
| ***Allah*** | ***(god)*** | ***The name of God among Muslims*** | ***الله (إلـه)*** |  |
| ***Allah*** | **name** | Allah's Best Names and His Glory's Characteristics | **اسم** | اسماء الله الحسنى وصفاته |
| ***rational*** | - | Creatures that are human beings, Ghosts, or Angels. | **عاقل** | إنسان أو ملك أو جن |
| ***rational*** | **prophet** | Male person who is a Prophet or messenger | **نبي** | نبي او رسول |
| ***rational*** | **people** | A group of people who share the same characteristics. | **ناس** | قوم ـ جماعة ـ قبيلة ـ فئة ـ مجموعة من البشر |
| ***rational*** | **person** | An individual or human being. | **إنسان** | فرد ـ شخص من البشر |
| ***rational*** | **ghost** | An invisible creature, created by Allah out of a "mixture of fire", who roamed the earth before human beings. | **جني** | فرد من الجن |
| ***rational*** | **ghosts** | A group of Ghosts. | **مجموعة** | جماعة من الجن |
| ***rational*** | **angels** | Celestial beings created from a luminous origin by Allah to | **ملائكة** |  |

| Class | Fine Class | Definition | صنف | تعريف |
|---|---|---|---|---|
| | | execute specific tasks he has given them. | | |
| ***rational*** | **angel** | One of the Angels. | **ملك** | |
| ***rational*** | **title** | Title of an Angel, ghost or person. | **لقب** | لقب لإنسان أو جان أو ملك |
| ***rational*** | **description** | Description of an Angel, ghost or person. | **وصف** | وصف او تعريف بشخص او ملك او جن |
| ***description*** | | ***description of abstract concepts*** | ***وصف*** | |
| ***description*** | **definition** | Definition of something. | **تعريف** | تعريف شيء ما |
| ***description*** | **description** | Description of something. | **شرح** | شرح أو تفسير |
| ***description*** | **manner** | Manner of an action: approach, course, manner, method, procedure, way, action, work. | **منهج** | سلوك - منهج - طريقة - فعل - عمل |
| ***description*** | **law** | An Islamic law or rule | **حكم** | حكم شرعي |
| ***description*** | **moral** | Concerned with the principles of right and wrong behaviour, and the goodness or badness of human character. | **خلق** | خلق صالح - او خصلة محمودة |
| ***description*** | **meaning** | Words or phrase meaning | **معنى** | معنى كلمة او تعبير |
| ***description*** | **other** | Other description | **آخر** | وصف لشيء آخر |
| ***description*** | **condition** | Conditions or qualifications. | **شرط** | شرط أو متطلبات للحصول على أمر ما |
| ***description*** | **speech** | Speech, supplications, dialogue or talk. | **قول** | قول - دعاء - إجابة - كلام |
| ***description*** | **attribute** | A distinctive attribute or aspect of something: feature, attribute, and characteristic. | **صفة** | حال- صفة - ميزة - خصلة - منقبة |
| ***description*** | **difference** | A point or way in which people or things are dissimilar. | **فرق** | فرق بين شيئين او معنيين |
| ***description*** | **result** | Result, consequence, outcome, effect. | **نتيجة** | نتيجة او محصلة لأمر ما |
| ***description*** | **request** | request - order | **طلب - أمر** | |
| ***description*** | **craft** | Craft, job or skill | **حرفة** | حرفة يدوية أو مهنة أو وظيفة |
| ***description*** | **proof** | Evidence, proof. | | **إثبات - دليل - برهان** |
| ***description*** | **favour** | God's bounty, grace and favour | **فضل** | فضل - نعمة - منة من الله |

| Class | Fine Class | Definition | صنف | تعريف |
|---|---|---|---|---|
| *description* | **list** | List, classes, types, stages. | **قائمة** | قائمة ـ صنف ـ نوع ـ مرحلة |
| *description* | **reason** | Reasons, cause, goal, aim. | **سبب** | سبب ـ هدف ـ المسبب لشيء ما |
| *location* | - | *locations* | **موقع** | |
| *location* | **afterlife location** | Mountains, rivers, valleys, heavens, hell fire. | **موقع في الاخرة** | موقع في الحياة الاخرة |
| *location* | **geographical location** | | **موقع جغرافي** | موقع على الارض |
| *location* | **city** | Cities, villages, countries. | | **مدينة ـ قرية ـ بلد** |
| *location* | **mountain** | Mountains. | **جبل** | جبل ـ تل ـ قمة |
| *location* | **other** | Other locations. | **آخر** | مكان آ خر |
| *location* | **place-of-worship** | A building used for public worship such as Mosque or church. | **دار العبادة** | مسجد أو معبد |
| *location* | **house** | Any type of houses | **بيت** | بيت ـ منزل ـ مسكن |
| ***number*** | - | **numeric values** | **رقم** | |
| ***number*** | **occurrence** | How many times something occurs. | **التكرار** | تكرار الكلمات في القران |
| ***number*** | **count** | Number of something. | **عدد** | عدد شيء ماء |
| ***number*** | **date** | Day of the month or year as specified by a number. | **تاريخ** | |
| ***number*** | **distance** | Linear measures. | **مسافة** | مسافة بين نقطتين |
| ***number*** | **value** | Estimate the monetary worth of (something). | **سعر** | قيمة ـ سعر |
| ***number*** | **order** | Order or rank. | **ترتيب** | |
| ***number*** | **other** | Other numbers. | **اخرى** | |
| ***number*** | **period** | The time something takes or lasts. | **فترة** | **فترة ـ مدة** |
| ***number*** | **percent** | Fractions. | **نسبه مئوية** | |
| ***number*** | **speed** | Speed or velocity. | **سرعة** | |
| ***number*** | **temp** | Temperature. | **درجة حرارة** | |
| ***number*** | **size** | Size, area and volume. | **حجم** | |
| ***number*** | **weight** | Weight. | **وزن** | |
| *event* | - | *a thing that happens* | **حدث** | |

| Class | Fine Class | Definition | صنف | تعريف |
|---|---|---|---|---|
| *event* | **calendar event** | Event occurs at a particular time every year. | **التقويم** | حدث على مدار السنة مثل الحج و العيد |
| *event* | **historic event** | Event happened in the past | **حدث تاريخي** | حدث في الماضي |
| *event* | **physical event** | Event happens in a particular time | **حدث حركي** | حدث مرتبط بوقت كصلاة الفجر |
| *event* | **afterlife event** | Event occurs in the afterlife. | **حدث الاخرة** | حدث يقع في الاخرة |
| *event* | **other** | Other event. | **اخرى** | |
| ***Polar question*** | | ***questions with a yes or no answer*** | ***تقريري*** | ***الأسئلة التقريرية او التخييرية*** |
| ***Polar question*** | **yes-no** | Agree or disagree. | **نعم-لا** | تأكيد أو نفي |
| | **choice** | Choosing between two or more possibilities. | **اختياري** | خيارات متعددة |

## 5.4 Arabic Question Words

An interrogative particle or a question word is generally used to ask for information. Just like in English, there are many different ways to ask questions in Arabic. The question particle usually occurs at the beginning of a sentence. First, here is an overview of the different types of questions that can be asked in Arabic.

All question particles in the Arabic language are classified as either ‘interrogative particles’ or ‘interrogative nouns’; both are called ‘أدوات الاســـتـفـهـام’- 'adawat alaistifham’ . The interrogative particles are أ and هل while the interrogative nouns are من , مـا , متى , أين , أيـان , أنى , كيف , كم ,  أي. These interrogative particles are sometimes connected by prepositions such as من بـ عن في لـ إلى على (Al-Fadili, 1980). Table 20 demonstrates some examples of Arabic question particles in detail. The Arabic question particles list is explained in Appendix D.

**Table 20: the description of Some Arabic Question Particles**

| الاستفهام عن<br>Ask about | English translation | أسماء الاستفهام<br>Question particles | الحروف المقترنة بها<br>Connected Prepositions | عبارة الاستفهام<br>prepositional phrase | | | | |
|---|---|---|---|---|---|---|---|---|
| المفعولية<br>Entity | Whom<br>للعاقل | من<br>(men) | (bi)بـ<br>(Ann)عن<br>(fi)في<br>(li) لـ<br>(ila)إلى<br>With – about in - for - to | بمن<br>With whom | عن من<br>About whom | فيمن | لمن<br>Whose | إلى من<br>to whom |
| | Which<br>عامة | أيّ<br>(ayu) | | بأيّ<br>with which | عن أيّ<br>About which | في أيّ<br>In which | لأيّ<br>For which | إلى أيّ<br>To which |
| | What<br>لغير العاقل | ما<br>(ma) | | بما<br>With what | عنما<br>About what | فيما<br>In what | لما<br>For what | إلى ما<br>To what |

The different types of questions that can be asked in Arabic can be categorised into: polar questions, wh-questions, and command questions.

A polar question[91] (plural, polar questions) can be defined as “A question which has only two possible responses: yes (affirmative) or no (negative)”. In the Arabic language, هل (hal) and أ (a) are both used in a question that demands a ‘yes’ or ‘no’ answer. Additionally, these particles could be used in multiple choice questions. For example,

أزرتم مكة أم المدينة المنورة؟

‘'Azurtum makkat 'amm  almadinat almunwrh?’

Did you visit Makkah or Madinah?

In English, wh-questions[92] are questions that are used to request information. The wh- question words are: who, whom, whose, what, where, when, why, which and, although it does not start with a wh-,  how.  The Arabic question words that are similar to the wh- English question words are: من , ما , متى , أين , أيان , أنى , كيف , كم and أي. Arabic wh-questions are usually used to ask about: entity, number, quantity,

[91] *https://en.wiktionary.org/wiki/polar_question*

[92] *https://dictionary.cambridge.org/dictionary/english/wh-question*

place, feeling, direction, time, person, reason, and things. For example, a question starting with مَن ('men') usually asks about a person.

Command questions typically begin with the imperative verb (verb 3) and ask for explanation, compression, opinion or evaluation. In English, for example, the command questions begin with: explain, summarise, describe, discuss, illustrate, define, analyse, compare, evaluate, assess, criticise, prove, support or justify. The Arabic command questions start with the imperative verb (verb 3), including: اشرح

, عدد, حدد, فسر, علل, وضّح, اذكر, بيّن, and, فسّر.

## 5.5 The Quran question classifier

The Quran question classifier is a new tool that can predict the answer type for any Arabic question about the Quran. This tool is built based on the deep learning technique called fastText. This section will present the existing Arabic question classification methods. After that, it will describe the fastText model in term of question classification. Finally, the section will discuss in detail the development of the new Arabic Quranic question classifier.

### 5.5.1 Existing Arabic question classification methods

Abdelnasser et al. (2014) used the Support Vector Machine classifier to classify Quranic questions into five types: creation, entity, location, number, physical and description. To train this SVM model, 230 questions were used that had been collected from different Islamic forums. Each question was tagged with one anticipated type. This dataset was divided into two groups: 180 questions used for training and 50 questions for testing. The accuracy of the classifier on this testing set was 86 percent.

Al Chalabi, Ray, and Shaalan (2015) classified Arabic questions by applying the context free grammar technique and the regular expressions technique. In this experiment, Nooj Platform[93] was employed to generate regular expressions and linguistic patterns are used to identify the class of the expected answer.

[93] *http://www.nooj4nlp.net/*

Hasan and Zakaria (2016) devised a question classification method that uses SVM (support vector machines). This technique categorises only questions that start with “Who”, “What”, and “Where”. The researchers applied one-gram, bi-gram, third-gram features and Term Frequency Weighting. They found that the best classification accuracy (87.25 percent when using the F1 metric) was obtained by using the bi-gram feature.

Waheeb and Babu (2016) proposed a question classification technique using SVM and Multinomial Naive Bayes (MNB) for Arabic questions. MNB is an advanced version of Naive Bayes that is designed for classifying text in documents. MNB uses word counts in a document rather than particular word presence and absence as the Naive Bayes classifier does. In this proposed model, the classifiers were trained on 300 Arabic questions taken from Wikipedia. The classifiers were then tested against 200 TREC questions translated from English into Arabic. The obtained F1-measure was 95 percent. The obtained F1-score metric by SVM was 97 percent.

All the proposed Arabic question classifiers mentioned above obtained more than 80 percent question classification accuracy. These results were obtained because these classifiers were trained on small training datasets with a few general question classes (not more than 6 classes such as Location and Human). These limitations might increase the ambiguity of the Quran question classification and lead to confusion; for example, by classifying God, ghost, Angel and Human under the HUMAN class.

To overcome this limitation and increase the predictive accuracy of classification models, the size of the training dataset must be increased ( Li & Roth, 1998). This would also increase the performance of the hierarchical classifier and help resolve ambiguities in the Arabic Quranic question classification.

### 5.5.2 The FastText model

The FastText[94] is a python library used for learning text classification and word embedding. It comprises an unsupervised learning algorithm based on character

[94] *https://fasttext.cc/*

n-grams to obtain vector representations for words. Facebook research centre developed a FastText[95] tool which classifies text using a supervised as well as an unsupervised learning algorithm. The Facebook research centre uses a labelled dataset which contains approximately 15,000 labelled questions with around 734 tags about cooking to train and test the FastText classifier. This dataset is split into 12,404 questions that are used as a training set and 3,000 questions that are used as a test set. In this experiment, the precision of predicting question labels ranges from 12.4 percent to 59.9 percent. The variation in the precision metric results is due to three factors: pre-processing the input data, changing the values of the parameters: epochs (range [5 - 50]), learning rate (range [0.1 - 1.0]), and word n-grams (range [1 - 5]).

### 5.5.3 Developing Quran question classifier

The fastText technique is used to develop the Arabic Quran questions classifier. In the Quran question classifier model, the fastText model used a dataset of 2200 labelled questions with 66 Quranic questions classes[96] from the AQQAC. These classes are the correct answer types of questions. Therefore, each question in this datasets is tagged with right class (answer type). For example, 'prophet' is the class of the question:

Who was sent to Thamood group ? lable__prophet__

These labelled questions are randomly split into 1800 questions that are used as a training set, and 400 questions that are used as a test set. This model has two main functions: training function and test function. The training function takes training set to create a model which can predict the question class. Then, the test function evaluates this model by computing the precision and recall at k. the k is the number of the top predicted classes.

[95] *https://github.com/facebookresearch/fastText/blob/master/docs/supervised-tutorial.md*

[96] Quranic questions classes are described in section 5.3.6

Before training this model, the input dataset is pre-processed; that is all Arabic diacritics, other symbols, and non-alphabetic characters are removed. For example:

| | |
|---|---|
| Before text -processing | على من تتحدث هذه الآية الكريمة: أَلَمْ تَرَ إِلَى الَّذِينَ يُزَكُّونَ<br>أَنفُسَهُمْ بَلِ اللّهُ يُزَكِّي مَن يَشَاءُ وَلاَ يُظْلَمُونَ فَتِيلاً{4}<br>__label__عاقل-ناس |
| After- text processing | __label__عاقل-ناس<br>على من تتحدث هذه الآية الكريمة ألمْ تر إِلى الذِين يزكون<br>أنفسهمْ بل الله يزكي من يشاء ولا يظْلمون فتيلا |

After the pre-processing stage, about 30 experiments were carried out to train and test the Quran question classifier model until the best accuracy of predicting a question class is obtained (the model python code is in Appendix D.5).

In each experiment, this classifier is trained on the same training-dataset including new values of the three parameters: thread, learning rate, and minimum n-gram. Then, this trained model is tested against the test-dataset by computing the precision and recall at k of top predicted class. the k in this case equals one.

This process was repeated until the best values relating to precision and recall metrics of the top predicted class were gained as can be seen in figure 21.

Following this, the same steps were carried out to achieve an acceptable level of accuracy of prediction for the best two question classes for each question in the test set file, as can be seen in figure 22. the precision and recall are computed at k=2 of top predicted class.

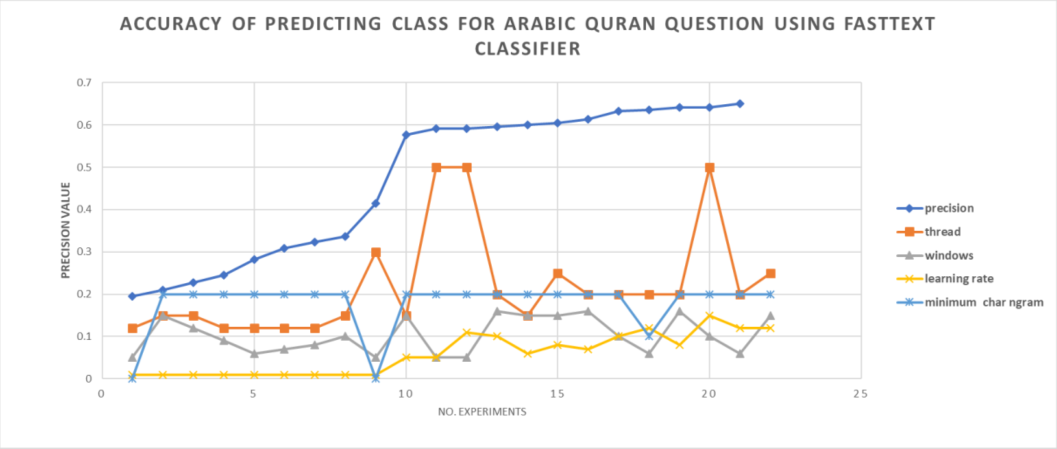


**Figure 21: Accuracy of predicting class for Arabic Quranic question using FastText classifier**

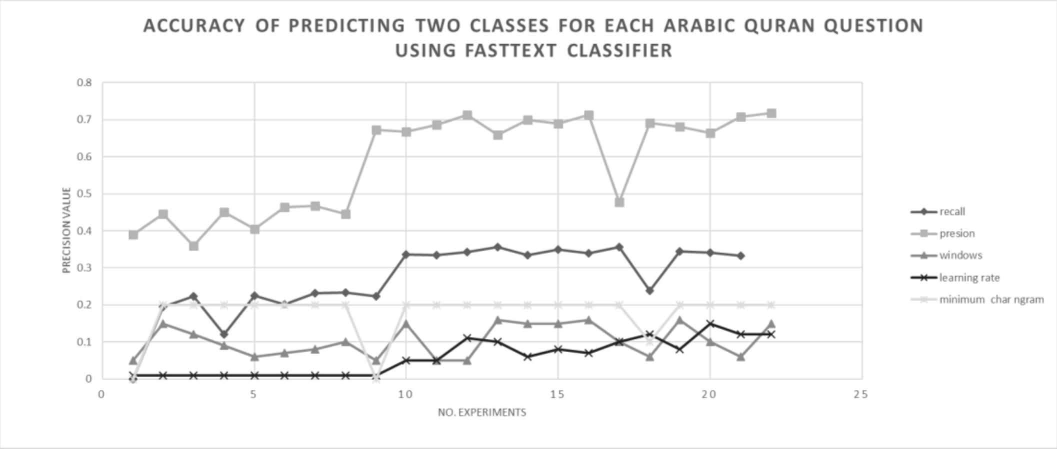


**Figure 22: Accuracy of predicting two classes for each Arabic Quranic question using FastText classifier**

## 5.5.4 Results

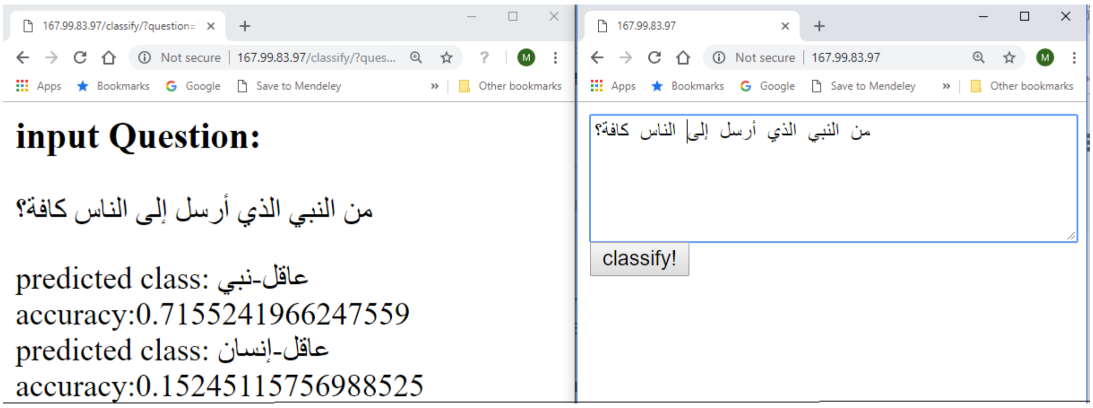
The results were very promising; the attained precision for predicting a question class by Arabic Quranic question classifier using FastText was 65.5 percent. Additionally, the recall for predicting the best two classes for a question was 71.8 percent. This accuracy of the Arabic Quranic question classifier could be enhanced by increasing the size of the training dataset for each question class.

### 5.5.5 Online Arabic Quranic question classifier tool (OAQQCT)

The Arabic Quranic question classifier tool[97] is an online Arabic classifier that can predict the top two answer types with result accuracy. This tool is based on the Arabic Quranic question classifier described in section 5.5.3. For instance, table 21 and figure 23 show the best top classes of an input question using the online classifier tool.

**Table 21: The classes of An Islamic question classified by OAQQCT**

| |
|---|
| من النبي الذي أرسل إلى الناس كافة؟<br>Who is the prophet was sent to all people? |
| **Result will be:** |
| predicted class: rational-prophet (عاقل-نبي)<br>accuracy: 72% |
| predicted class: rational-prophet (عاقل-إنسان)<br>accuracy:15.3% |


167.99.83.97/classify/?question=
Not secure | 167.99.83.97/classify/?ques...
Apps
Bookmarks
Google
Save to Mendeley
Other bookmarks
167.99.83.97
Not secure | 167.99.83.97
classify!

**input Question:**

من النبي الذي أرسل إلى الناس كافة؟

predicted class: عاقل-نبي
accuracy:0.7155241966247559
predicted class: عاقل-إنسان
accuracy:0.15245115756988525

من النبي الذي أرسل إلى الناس كافة؟

**Figure 23: The classification of an Arabic Islamic question by OAQQCT**

[97] http://167.99.83.97/

## 5.6 Semantic similarity model for the Quran Arabic words using word embeddings

### 5.6.1 Word embeddings

In linguistics, word embeddings can be defined as a process of quantifying and categorising semantic similarities between words depending on their distributional properties in huge samples of language data. Word embedding is a common term for a set of language modelling and feature learning techniques in natural language processing (NLP), where words or phrases from the vocabulary are mapped to numerical vectors (Lebret, 2016).

In NLP, word embedding can be used in syntax analysis, idiomaticity analysis, semantic analysis, sentiment analysis, Part Of Speech tagging, named entity recognition, machine translation, textual entailment, and word meanings ( Li & Yang, 2017).

### 5.6.2 Word embeddings techniques

Recent techniques using word embeddings are Word2vec (Mikolov, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), FastText (Joulin et al., 2016), Gensim (Rehurek & Sojka, 2010), and Deeplearning4j (Eclipse Deeplearning4j Development Team, 2017).

The word2vec tool takes a text corpus as input and produces the word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representation of words. The resulting word vector file can be used as the features in many natural language processing and machine learning applications.

There are two main learning algorithms in word2vec: continuous bag-of-words and continuous skip-gram. The switch-cbow allows the user to pick one of these learning algorithms. Both algorithms learn the representation of a word that is useful for prediction of other words in the sentence.

FastText[98] is a library for learning of word embeddings and text classification created by Facebook's AI Research lab. The model allows the creation of an

[98] *https://fasttext.cc/*

unsupervised learning or supervised learning algorithm to obtain vector representations for words.

Gensim[99] is an open source vector space and topic modelling toolkit implemented in Python. Gensim contains implementations of word2vec, FastText, and document2vec algorithms. The Gensim package is actually an extended version of the Google Word2Vec package with additional functionality.

Building a good data model for term similarity features involves using a deep-learning technique and training on a large dataset of plain text collected from different text sources belonging to the same knowledge domain, such as the Quran knowledge domain. An example of a training dataset is the Google News dataset[100]that contains about 100 billion words. The Facebook Research Centre also distributes pre-trained word vectors for 157 languages[101]. These models are trained on Common Crawl and Wikipedia using FastText.

### 5.6.3 Developing a similarity model for the Quran Arabic words using Genism

To create a similarity model for the Quran Arabic words, an Islamic Arabic dataset (IAD) is needed to train the word2vec algorithm. The IAD was developed from a collection of 1061 classical Arabic language books which are mainly related to the Quran sciences. This dataset consists of more than 150 million words. Most of the books were collected by Alsheddi (2016).

The Islamic Arabic dataset went through text pre-processing to remove non-Arabic letters and diacritics ‘التشـكيل’ (vowel marks). Following this, the texts of the books were gathered into one text file in which each line has only one sentence. Finally, each sentence was tokenised using space tokenisation.

The next step was loading the IAD to the word2vec algorithm to create vocabulary which is a set of unique words. Once this was done, the word2vec model training could commence. The result of this training is learned vectors which are features

[99] *https://radimrehurek.com/gensim/*

[100] *https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing*

[101] *https://fasttext.cc/docs/en/crawl-vectors.html*

that describe each word in a vocabulary. The vocabulary size in this model is around 700,000 words.

Figure 24 shows the list of words that have a similar semantic meaning to given words.

| Most similar words for ['غــزوة','بــدر' ghazawht Bader] | Similar word for [ 'عيسى' Essa ] |
|---|---|
| Results:<br>[ ('0.666696310043335 ,'الخندق),<br>('0.6656983494758606 ,'تبوك),<br>('0.6562150716781616 ,'الحديبية),<br>('0.6467557549476624 ,'خيبر),<br>('0.6378562450408936 ,'مؤتة),<br>('0.6210472583770752 ,'سرية),<br>('0.5845927000045776 ,'اليرموك),<br>('0.5836865305900574 ,'أوطاس),<br>('0.5780510902404785 ,'وقعة),<br>('0.5632182359695435 ,'صفين),<br>('0.5627709627151489 ,'القادسية),<br>('0.5592902898788452 ,'حنين), | Results:<br>[('0.6712440252304077 ,'وعيسى),<br>('0.6557479500770569 ,'محمد),<br>('0.6078304052352905 ,'موسى),<br>('0.6038110256195068 ,'إبراهيم),<br>('0.5851383209228516 ,'هارون),<br>('0.5793220400810242 ,'يوسف),<br>('0.5736751556396484 ,'صالح),<br>('0.5646899938583374 ,'ابراهيم),<br>('0.5563672780990601 ,'يعقوب),<br>('0.5562412142753601 ,'إسماعيل)] |

**Figure 24: Example of finding similar words for a given word**

## 5.7 Conclusion

This chapter provided some background to text corpus linguistics and Question and Answer (Q&A) corpora. The chapter emphasised the importance of Q&A corpora and summarised their use in Information Retrievals Systems such as in question classification and evaluation of IR systems. It then described the construction of the Arabic Quranic Question and Answer Corpus (AQQAC) in some detail. The AQQAC is a collection of around 2,224 questions and answers about the Quran. Each pair of question and answer is annotated with the question ID, the question word (particles), the chapter number, the verse number, the question topic, the question type and the Quran ontology concepts. This AQQAC can be accessed via *https://doi.org/10.5518/356.* Finally, this chapter described the classification of the Arabic Quran Q&A using the Quran ontology. The Quran question taxonomy uses a two-layered question hierarchy which contains 8coarse classes and 85 fine classes. Finally, the chapter explained the new learning Arabic Quranic Question Classifier with an average accuracy of 72%

# Chapter 6
# Semantic Arabic Quranic search model based on the Quran ontology

This chapter will explain a new Arabic Quranic semantic search model (AQSSM) for the Quran based on the Quran ontologies (QO) developed in Chapter 4. This model attempts to retrieve Quranic verses to answer a question about the Quran in the Arabic language. This model also proposes a new technique that attempts to overcome the limitations of the Quranic search applications reviewed in Chapter 3.

The work presented in this chapter has appeared in the following publications:

Alqahtani, M.M.A., Atwell, E. (2015). A Review of Semantic Search Methods to Retrieve Information from the Qur'an Corpus. In 8th international Corpus Linguistics Conference, Lancaster, pp. 21–24, Jul 2015.

Alqahtani, M.M.A., Atwell, E. (2015). Qur'anic search tool based on ontology of concepts. In the 8th Saudi Students Conference in London, pp. 29–30, Jan 2015.

Alqahtani, M., & Atwell, E. S. (2016, June). 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. Vol. 9612, pp. 478–485: Springer International.

## 6.1 Introduction

Two types of Quran search models exist: semantic-based and keyword-based models. The semantic-based technique is a concept-based search tool that retrieves results based on word meaning or concept match, while the keyword-based technique returns results based on letter-matching. The majority of Quranic search tools use the keyword-based technique (Mohammad Alqahtani & Atwell, 2017).

The existing Quranic semantic search techniques include an ontology-based technique, a synonyms-set and a cross-language information retrieval (CLIR) technique. The ontology-based approach searches for concept(s) that match the words in the user's query and returns verses that are related to these concept(s).

The synonyms-set technique produces all synonyms of the words in the query using WordNet and finds all Quranic verses that match the synonyms for these words. CLIR translates the words in the query to another language and then retrieves verses that contain words matching the translated words.

Several deficiencies can be found in Quranic verses (Ayat) that are retrieved using the existing keyword search techniques. For example, the answer to the query might have irrelevant verses, the answer could skip relevant verses, and the retrieved verses might not be in the correct order.

The limitations of the keyword-based technique include: misunderstanding the exact meaning of the words in the query and neglecting theories of information retrieval (Raza et al., 2014). Current Quranic semantic search techniques also have limitations in finding requested information because they use incomplete QO. Moreover, these ontologies do not cover all aspects and facts in the Quran and were developed to meet semantic web standards (Alqahtani & Atwell, 2018; Alrehaili & Atwell, 2014).

This chapter is organized as follows: Section 6.2 discusses the framework of Arabic Quranic semantic search tools; Section 6.3 describes the methodology of the Arabic Quranic Search model based on ontology; and Section 6.4 summarizes the critical conclusions of this chapter.

## 6.2 Framework of Arabic Quranic semantic search model

A new Arabic Quranic Semantic Search Model (AQSSM) that is based on ontology aims to employ IR techniques and semantic search technologies. This tool was designed by combining the theories of previous research (Al-Yahya, Al-Khalifa, Bahanshal, Al-Odah, & Al-Helwah, 2010; Mony, Rao, & Potey, 2014; Ou et al., 2008; Yauri, 2014) and is divided into six components: QO, Quranic database (QDB), natural language analyser (NLA), SSM, keyword search model (KSM), and scoring and ranking model (SRM). Figure 25 demonstrates the components of the framework of AQSSM.

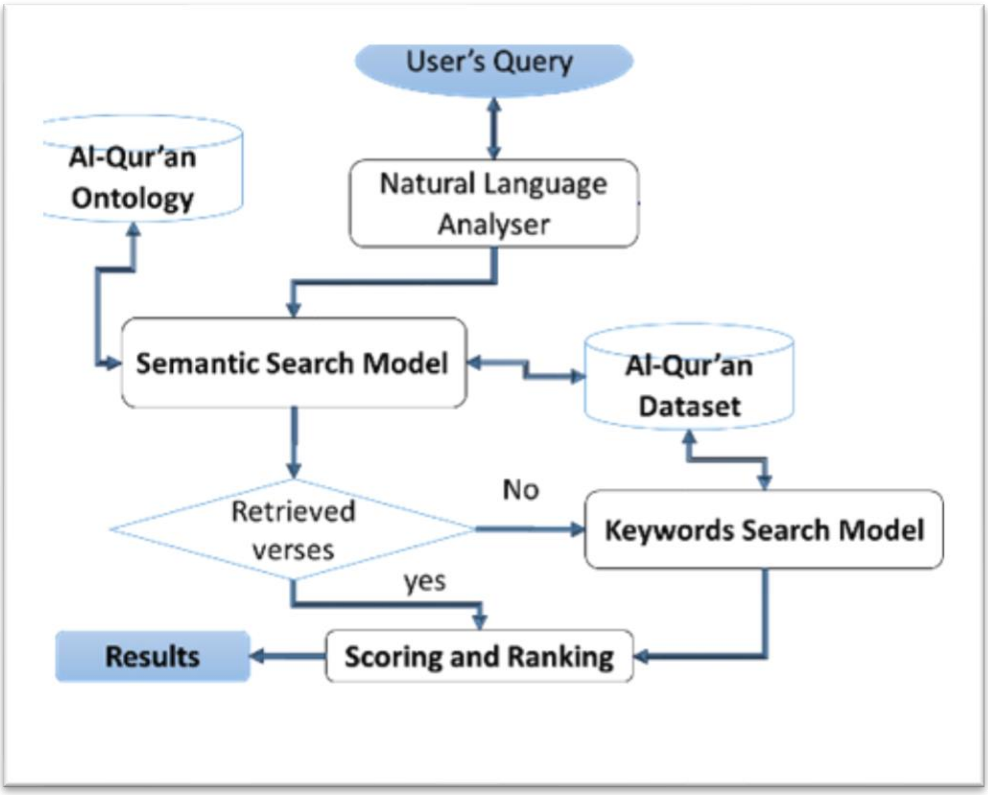


**Figure 25: Arabic Quranic Semantic Search Model structure (AQSSM)**

The QO component includes the newly developed QO which was discussed in Chapter 4. The QDB component consists of the Quran text in the Arabic language and the English translation of the Quran, as well as two different *Tafsirs* (description of the Quran), the Quran words dictionary, and concepts found in the Quran.

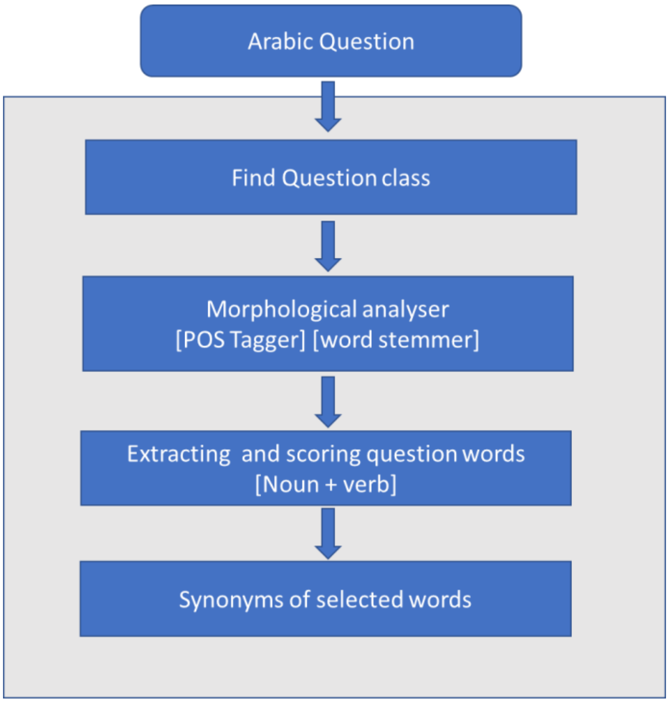

Arabic Question
Find Question class
Morphological analyser
[POS Tagger] [word stemmer]
Extracting and scoring question words
[Noun + verb]
Synonyms of selected words

**Figure 26 : Arabic question analyser**

The user query in the NLA model undergoes several steps as shown in figure 26: classifying the question, morphological analysis, extraction and scoring question words, and semantic meaning of extracted question's words.

The first step, the NLA finds the user's question class (answer type) using the Arabic Quranic classifier (previously described in section 5.5.3).

The second step, the NLA applies different NLP techniques, including part of speech (POS) tagging and Arabic stemmer. This study uses the Stanford CoreNLP[102] (Manning et al., 2014) to find the POS tag for each word, and ISRI Arabic Stemmer[103] from Natural Language Toolkit (NLTK). In this step, all nouns and verbs in the query are selected based on the POS tag of each word.

Next, NLA uses the Arabic similarity Quranic words model (previously explained in section 5.6) to generate synonyms for the verbs and nouns in the query.

[102] *https://stanfordnlp.github.io/CoreNLP/*

[103] *NLTK is a platform for building Python programs to work with natural languages. NLTK provides text processing libraries for classification, tokenization, tagging, stemming, and semantic reasoning. https://www.nltk.org/_modules/nltk/stem/isri.html*

The Final step, NLA adds the semantic tags generated during the previous steps to these words as shown in table 22 and then sends the results to SSM.

**Table 22: Example of analysing the words in a query**

| | |
|---|---|
| **Question in Arabic language**<br>**(transliteration)** | من هو النبي الذي أرسل إلى قوم ثمود؟<br>man hua annabiu aladhi 'ursil 'iilaa qawmi thamudan? |
| **Question in English language** | Who is the prophet who was sent to the people of Thamud? |
| **Question class (semantic tag)** | عاقل-نبي<br>Rational: prophet |
| **Part of Speech tags and weights for the words in the question** | ('من', 'WP': Wh-pronoun),0<br>('هو', 'PRP': Personal pronoun), 0<br>('الـنــبــي', 'DTNN': noun singular with the determiner),2<br>('الذي', 'WP': Wh-pronoun), 0<br>('أرسل', 'VBN': verb),1<br>('إلى', 'IN': preposition),0<br>('قوم', 'NN': noun), 2<br>('ثمود', 'NN': noun),2<br>('؟', 'PUNC'): punctuation), 0 |
| **Similar words to 'أرسل'**<br>**Similar words to 'نبي'** | بعث، أتي، دعا<br>رسول |

SSM searches the QO triples using the cosine similarity of vector space models (SVM) to find the most relevant concepts to the normalised query and then returns the result to SRM. However, if no result is found, KSM searches for verses that contain words that match the query.

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 and whether or not the ontology concepts match the question and answer. Finally, SRM provides the results to the user and records the selected result.

Table 23 presents the results when a user searches for 'What are the names of "the Paradise" in the Quran'?

ماهي أسماء الجنة؟

**Table 23: Answers to a question about names for 'the paradise'**

| Match ontology concepts | Number of retrieved verses | Average match percentage | Final decision |
|---|---|---|---|
| أسماء الجنة – Names for 'the paradise' (23 different names of the paradise) | 54 different verses | 81 percent | Returned as an answer to the question |
| الجنة | One verse | 70.7 percent | |
| صفات الجنة - the Paradise Features | 240 different verses | 63 percent | Suggested as a related topic |
| أصحاب الجنة – Companions to 'the paradise' | 33 different verses | 61 percent | Suggested as a related topic |

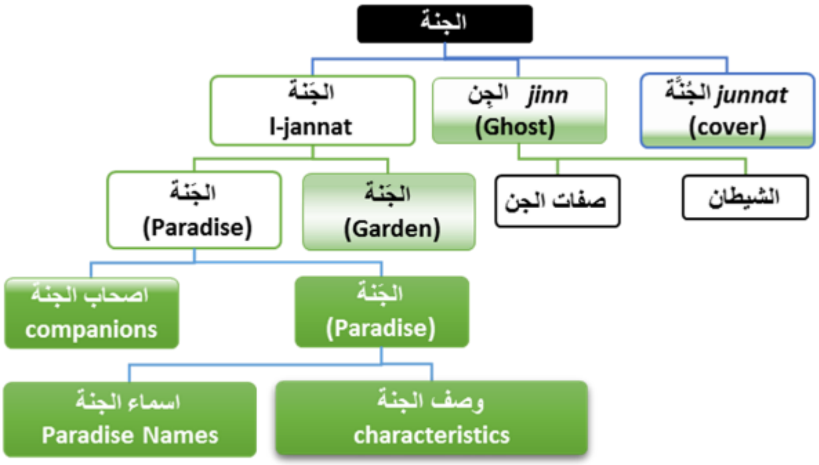


**Figure 27 : Search results for أسماء الجنة in AQSSM (darker colour represents a more relevant result).**

## 6.3 Model evaluation

### 6.3.1 The IR systems evaluation metrics for AQSSM

Evaluation metrics in IR systems aim to evaluate how well search results match the intent of the query. The most popular evaluation measures used for IR systems are: recall, precision, F-measure, precision at k, recall at k and average precision(previously described in section 2.2.3).The recall, precision and F-measure are used to measure the effectiveness of the unordered retrieved set (Zhu, 2004). Consequently, these metrics could not be appropriate to evaluate the ranked retrieved results when AQSSM was tested by a set of Quranic questions.

However, precision at k, recall at k and average precision could be used to assess the ordered retrieved documents (Manning et al., 2008). Subsequently, these metrics could be suitable to assess the ranked results of testing AQSSM by a set of Quranic questions.

Precision at k is the percentage of recommended items in the top-k set that are relevant. Using precision and recall at *k* to rank document retrieval make the process of computing and interpreting the metrics simpler. However, the drawbacks are that the value of k has a massive effect on the metric, and any ranking above *k* is inconsequential. In other words, no rule exists to determine the best value of k. Therefore, to overcome the issue of choosing the k-value to calculate precision, the average precision metric can be used to calculate precision and recall without having to set k value.

This average precision can be used for a single query. The mean average precision (MAP) is used to calculate the average precision of a set of questions. the MAP is computed by the following formula:

$$MAP = \frac{\sum_{q=1}^{\#Queries} Average\_Precision(q)}{\#\ Queries}$$

According to the functions of ranked evaluation metrics, MAP is the appropriate measure to evaluate the ranked results of testing AQSSM by a set of Quranic questions. For example, tables 24–26 demonstrate different evaluation results by the IR evaluation metrics for the retrieved verses to the question: ‘What are the names of “the Paradise” in the Quran?’ (the retrieved results were shown in Table 22). These IR evaluation metrics were recall, precision, F-measure, rated recall, rated precision, and the MAP.

**Table 24: Recall, precision and F-measure for retrieved verses containing information about names for ‘the paradise’.**

| |
|---|
| Recall = 54/54 = 100 % |
| Precision = 54/55 = 98.2% |
| F1= 2(100 * 98.2)/(100+98.2) = 99% |

**Table 25: The recall and precision @ 5 for retrieved verses that contain information about names for ‘the paradise’.**

| Rank (5) | Is relevant | Recall @ 5 | Precision @ 5 |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 1 | 0.018 | 0.5 |
| 3 | 1 | 0.037 | 0.667 |
| 4 | 1 | 0.055 | 0.75 |
| 5 | 1 | 0.074 | 0.8 |
| average | | 0.037 | 0.543 |

**Table 26: The recall and precision @ 10 for retrieved verses that contain information about names for ‘the paradise’.**

| Rank (10) | Is relevant | Recall | Precision |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 1 | 0.0185 | 0.5 |
| 3 | 1 | 0.037 | 0.667 |
| 4 | 1 | 0.055 | 0.750 |
| 5 | 1 | 0.074074 | 0.80 |
| 6 | 1 | 0.093 | 0.833 |
| 7 | 1 | 0.111 | 0.857 |
| 8 | 1 | 0.130 | 0.875 |
| 9 | 1 | 0.148 | 0.889 |
| 10 | 1 | 0.167 | 0.90 |
| Average | | 0.083 | 0.707 |

### 6.3.2 Experiments and results

Two types of experiments were conducted to measure the performance and accuracy of this model. The first experiment aimed to measure the accuracy of the results retrieved by the keyword search model. The second experiment aimed to measure the answers accuracy of the AQSSM, which combines the keyword and semantic models.

The testing questions set consists of one hundred questions. These questions were selected randomly from the Arabic Quranic question and answer (previously described in section 5.3). Each question in this set was linked to the right answer that is Quranic verses. The testing set was used in the two experiments.

In both experiments, the testing set was sent to the both models. Then, the models returned the answer for each question as a list of ranked verses from the most to the least relevant to the answers. Then these results were compared to the right answers in the testing set.

Experiment 1 shows that the keyword search model answered 24 out of 100 questions with an average recall rate of 15.1 percent and a MAP rate of 4.6 percent. Figure 28 presents the chart of the average precision level for each question answered using the keyword search model.

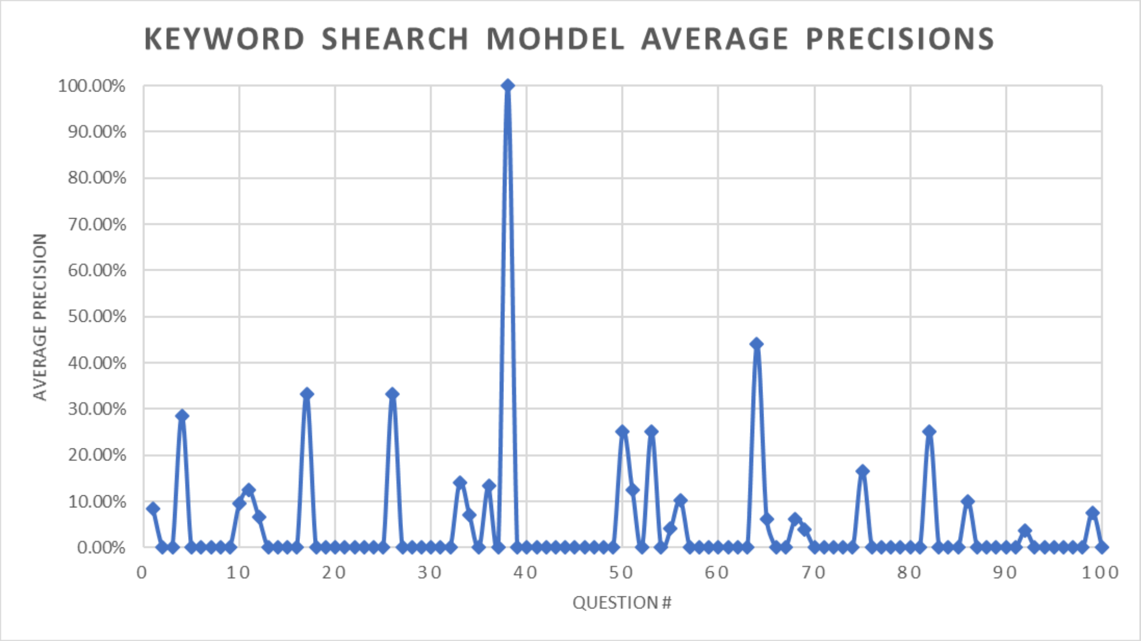


**Figure 28.** Average precision level for each question answered by the keyword search model.

Experiment 2 shows that the AQSSM answered 96 out 100 questions with an average recall rate of 41 percent and a MAP rate of 58.3 percent. Figure 29 shows the chart of the average precision level for each answer to a question using AQSSM.

By comparing the results of the two experiments, the keyword search model answered 24 percent of the questions with a recall rate of 15.1 percent and a MAP rate of 4.6 percent. In contrast, the AQSSM answered 96 percent of the questions with an average recall rate of 41 percent and a MAP rate of 58.3 percent.

It is clear that AQSSM answered the questions more accurately than the keyword search model. However, AQSSM fails to answer the questions asking about numbers and explanations. A possible explanation for this might be that the AQSSM was designed to return only Quranic verses as an answer to a question.

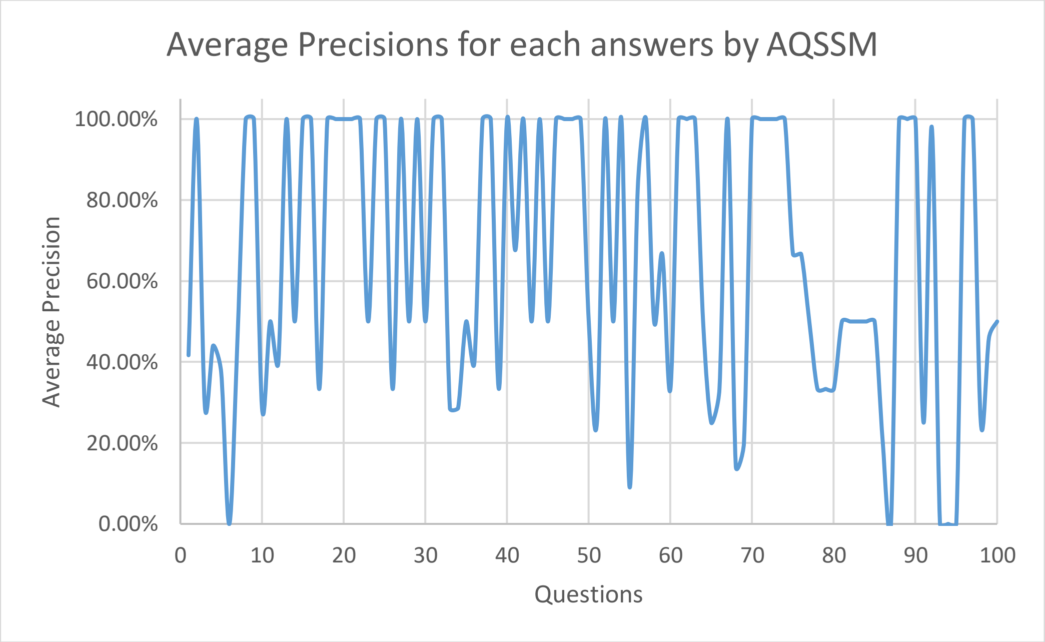


**Figure 29.** Average precision level for each question answered by AQSSM.

Table 27 presents some of the answers to questions about the Quran using AQSSM.

**Table 27: Some answers to questions about the Quran by AQSSM**

| Question | Chapter No. | Verse No. | Quran Concept | ranking score |
|---|---|---|---|---|
| ما هي أول آية في كتاب الله الخالد بعد البسملة؟ | 1 | 2 | آية أول آية | 0.4472 |
| سئل الكسائي كم آية في القرآن أولها شين؟ فأجاب :أربع آيات، فما هي؟ | 2 | 185 | آية آيات أولها شين | 0.6201 |
| الحمد لله رب العالمين (تكررت في القرآن الكريم كآية أو بعض آية ست مرات، في ست آيات، فما هي؟ | 1 | 2 | آية الحمد لله رب العالمين | 0.527 |
| (عن مقاتل والكلبي) :كل فحشاء في القرآن فهي الزنى إلا في هذا الموضـــع فإنها البخل.(ففي أيّ موضـــع وفي أيّ آية جاءت الفحشاء بمعنى البخل؟ | 2 | 268 | آية الفحشاء | 0.2886 |

| Question | Chapter No. | Verse No. | Quran Concept | ranking score |
|---|---|---|---|---|
| أشـــارت آية كريمة من آيات القرآن الكريم إلى وجوب الوقوف بعرفة والإفاضة منها، فما الآية الكريمة الدالة على ذلك؟ | 2 | 198 | آية وجوب الوقوف بعرفة والإفاضة منها | 0.6123 |
| اتفق العلماء على أنّ الغنيمة تقســم خمســة أقســام، فيعطى الخمس لمن ذكرهم الله في الآيـة، والباقي يوزع على الغـانمين، ويقســم الخمس خمســة أســهم، ســهم للرســول، وســهم لذوي القربى، وسـهم لليتامى، وسـهم للمسـاكين، وسـهم لابن السـبيل. فما الآية التي ورد فيها توزيع الغنائم | 8 | 41 | الخمس | 0.2562 |
| نهى الإســلام عن الرشــوة التي تؤدي إلى تعطيل الشريعة، وتملأ نفوس النـاس بـالأحقـاد، وتؤدي إلى تراكم الثروة في أيـدي فئـة بدون حق وإلى انعدام الثقة بين أبناء الأمة، فما الآيات القرآنية الكريمة الدالة على ذلك؟ | 2 | 188 | الرشوة | 0.2041 |
| قال تعالى :الزَّانِيَةُ وَالزَّانِي فَاجْلِدُوا كُلَّ واحِدٍ مِنْهُما مِائَةَ جَلْدَةٍ ]النور [2 :لماذا قدم الزانية على الزاني بينما في السرـقة قدم السارق على السارقة؟ | 24 | 2 | الزانية والزاني | 0.554 |
| ما عدة الزوجة إذا كانت مدخولا بها ومن ذوات الحيض؟ | 2 | 228 | الزوجة | 0.3535 |
| ما عدد السماوات؟ | 2 | 29 | السماوات | 0.7071 |
| حرم الإســلام كتمان الشــهادة، وقد جاء هذا التحريم في ثلاث آيات فما هي؟ | 2 | 283 | الشهادة | 0.2886 |
| اذكر الآيـة الـدالـة على أن الصــــدقـات تكفر بعض الـذنوب والسيئات. | 2 | 271 | الصدقات | 0.4082 |
| ماذا كانت دعوة امرأة عمران عندما أحســت بالحمل بعدما أسنت واشتاقت للولد؟ | 3 | 35 | امرأة عمران | 0.4264 |
| ما قصة النمرود في القرآن؟ | 2 | 258 | قصة النمرود | 0.8164 |
| ما قصة عزير في القرآن؟ | 2 | 259 | قصة عزير | 0.8164 |
| قال تعالى :وَلَقَدْ عَلِمُوا لَمَنِ اشْــتَراهُ ما لَهُ فِي الْآخِرَةِ مِنْ خَلاقٍ ]البقرة [102 :ما معنى قوله تعالى :خَلاقٍ؟ | 2 | 102 | معنى خلاق | 0.4160 |

Figure 29 demonstrates the average precision level for each question answered by AQSSM.

## 6.4 Conclusion

This chapter explained the framework of the Arabic Quranic semantic search model (AQSSM) for the Quran based on the Quran ontologies. This model retrieves Quranic verses as an answer to a question about the Quran in the Arabic language. This model also suggests hybrids techniques that attempt to overcome the limitations of the Quranic search applications reviewed in Chapter 3.

In general, this study showed that AQSSM considered most of these limitations. The evaluation of this model showed that AQSSM answered 96 per cent of questions with an average recall rate of 41 per cent and MAP rate of 58.3 per cent. However, AQSSM fails to answer the questions asking about numbers and explanations. A possible explanation for this might be that the AQSSM was designed to return only Quranic verses as an answer to a question. In contrast, this model succeeds to answer questions asking about entities. This may be explained by the fact that AQSSM matches the entity in the question with concepts in the Quranic verses using the Quranic ontology.

# Chapter 7
# Conclusion and Future Work

## 7.1 Summary of the work

The Quran is the main resource for the Islamic sciences and the Arabic language. Therefore, numerous Quranic search applications have been developed to facilitate the retrieval of knowledge from the Quran. The techniques used to retrieve information from the Quran can be classified into two types: semantic- and keyword-based techniques. Most Quranic search tools use the keyword search technique. However, many deficits occur in the retrieved Quranic verses for a query when keyword search techniques are utilised. Examples of these limitations are irrelevant retrieved verses, un-retrieved relevant verses or unranked retrieved verses. Additionally, keyword-based techniques may involve issues, such as misunderstanding the exact meaning of the input words forming a query and disregarding some theories of information retrieval. Current Quranic semantic search techniques also have some limitations in finding the needed information because these semantic searches use an inaccurate and uncompleted Quranic ontology.

This thesis aimed to present a detailed and novel methodology for developing a new Arabic Quran semantic search model. This methodology, which represents the main contribution of this thesis, includes a combination of Quran science resources, evaluation criteria for both Quran search tools and ontologies, and a model developed for an Arabic Quran semantic search tool. The resources include the Arabic–English Quran ontology, as well as the first Annotated Corpus of Arabic Questions and Answers on the Quran. Additionally, a Quran question taxonomy based on the Quran ontology.

This thesis reviews the majority of search tools constructed for the Holy Quran. Then, it evaluates these different search tools against 13 criteria in terms of search features, output features, precision of the retrieved verses, recall database size and types of database content. The comparison shows that the majority of Quranic search tools still cannot address ambiguity in the retrieved results.

After that, this thesis reviews previous Quranic ontologies and then compares them against 13 measures. Some deficits have been found in most of these ontologies, such as missing evaluation by Islamic scholars or applications. Additionally, a single Quranic ontology does not cover most of the knowledge in the Quran. Therefore, a new Arabic–English Quran ontology is developed from 10 datasets related to the Quran, such as the Quran chapter and verse names, Quran word meanings and Quran topics. The main aim of developing a Quranic ontology is to facilitate the retrieval of knowledge from the Quran. Quranic ontologies will also enrich raw Arabic and English Quran text with Islamic semantic tags.

Furthermore, this thesis developed the first Annotated Corpus of Arabic Questions and Answers on the Quran. This corpus has around 2,200 pairs of questions and answers collected from trusted Islamic books. Each pair of question and answer is labelled with five tags. Examples of tags are question type (either factoid or descriptive), topic of question based on the Quran ontology and question class. This corpus is used to evaluate the performance of the Arabic Quran semantic search model.

Moreover, the thesis introduced the Quran question classifier based on the annotated Quran question–answer corpus. This classifier was built based on fastText machine learning classification.

Finally, it explains a new search model, which is a semantic search model for the Quran based on Quran ontologies. This model aimed to overcome most limitations in existing Quranic search applications. This search tool uses both IR techniques and semantic search technologies.

## 7.2 Thesis achievements

The contributions of this research to the literature and its novelty are as follows:

1. Used 13 criteria to evaluate existing search tools that were constructed for searching the Quran
2. Used 14 measures to compare different Quran ontologies
3. Built a new Arabic–English Quran ontology from more than 10 datasets in the domain of the Quran. This ontology has more than 1,070,000 triples, 78 classes, 51 object properties, 34 data properties and 86,588 individuals.

4. Constructed the first Annotated Corpus of Arabic Questions and Answers on the Quran. This corpus is a collection of around 2,224 questions and answers on the Quran. Each pair of question and answer is annotated with a question ID, question word (question particles), Quranic chapter number, verse number, question topic, question type and the Quran ontology concept(s). The corpus is available online at https://doi.org/10.5518/356.
5. The Quran question taxonomy uses a two-layered question hierarchy which contains eight coarse classes (entity, God, rational, description, location, number, event, polar question) and 85 fine classes.
6. Developed a word embedding model for Arabic words in the Quran sciences domain. This model (called the Arabic similarity Quranic word model) was developed from a collection of 1,061 classical Arabic language books which are mainly related to the Quran sciences. The dataset consists of more than 150 million words. The vocabulary size of this model is around 700,000 words. This model receives a word and returns the most similar words with a similarity percentage.
7. Developed a new semantic search model for the Arabic Quran based on the new Quran ontology. This model accepts Arabic questions on the Quran. Then, it retrieves the most relevant Quran verses containing the answer. The evaluation results demonstrate that the overall accuracy of this model can achieve 58.3 %.
8. Developed the Learning Arabic Islamic Question Classifier with an average accuracy of 72%

## 7.3 Evaluation and future work

The evaluation of the work in this thesis and the future work that will be conducted for each aspect of the study as an extension are described in the following:

1. The new Quran ontology is constructed from 12 datasets related to the Quran. This ontology has more than 1,070,000 triples, 78 classes, 51 object properties, 34 data properties and 86,588 individuals. However, this ontology still does not cover all aspects of the Quran domain, such as Islamic ethics, principles, and laws

and rules of Muslim society. Developing hadith[104] ontology by using different hadith resources might help cover these Islamic aspects. This can be achieved by integrating hadith ontology with the Quran ontology.

2. The Arabic Quran Question-Answer corpus can be enhanced and expanded from 2,200 to at least 10,000 question and answer pairs. This corpus will include question–answer datasets from another Islamic resource, such as the Islamic hadith and fqih (Islamic law). This extended question–answer corpus will help research conducted in the Islamic field by using this dataset as a gold standard training and testing dataset to enhance the accuracy of Islamic information retrieval systems. Additionally, this dataset can be used in another question and answer research, such as those on matching duplicated questions and categorising questions in an appropriate group and training dataset for Arabic question classifiers.

3. The Learning Arabic Islamic Question Classifier has an average classification accuracy of 72%. This classifier can be enhanced by increasing the size of the training dataset of the Islamic questions and answers on the Quran from 1,800 questions to be at least 10,000 questions. The new Arabic question classifier could be used to classify not only questions from the Quran domain but also questions about the Islamic sciences.

## 7.4 Limitations

Despite the achievements of this project, there are two limitations which do not affect directly the project outcome.

Firstly, The Annotated Arabic Quranic question and answer corpus (AAQQAC) was not reviewed and approved by an Islamic scholar due to the project time limit. Additionally, the tag set of question classes should be review and approved by the

[104] *The term hadith refers to reports on the statements or actions of Muhammad, or of his tacit approval or criticism of something said or done in his presence (Campo, 2009). The Oxford Dictionary defines a hadith as 'a collection of traditions containing sayings of the prophet Muhammad which, with accounts of his daily practice (the Sunna), constitute the major source of guidance for Muslims apart from the Koran'.*

Islamic scholar. This limitation might affect the accuracy of the Arabic Quranic question classifier as this classifier was trained on AAQQAC.

The AAQQAC can be enhanced and expanded from 2,200 to at least 10,000 question and answer pairs. This corpus will include question–answer datasets from another Islamic resource, such as the Islamic hadith and fqih (Islamic law). This extended question–answer corpus will help research conducted in the Islamic field by using this dataset as a gold standard training and testing dataset to enhance the accuracy of Islamic information retrieval systems. Additionally, this dataset can be used in another question and answer research, such as those on matching duplicated questions and categorising questions in an appropriate group and training dataset for Arabic question classifiers.

Secondly, The Arabic semantic similar Quranic word model could retrieve similar words for a given word with a similarity of 75%. However, this model still has some drawbacks in the retrieved similar words, such as those words shown in Table 28 for القران‘the Quran’:

**Table 28: Arabic semantic similar words to القران ‘the Quran’**

| **Arabic semantic similar word to the Quran:** | |
|---|---|
| Arabic word: English translation | Arabic word transliteration |
| 'الكتاب': the book | ‘alkitab’ |
| 'التلاوة': the recitation | ‘altlawt’ |
| 'التنزيل': the revelation | ‘alitnzil’ |
| 'المصحف': the Quran | ‘almshuf’ |
| 'التوراة': the Torah | ‘alitwrat’ |
| 'الوحي': the revelation | ‘alwhy’ |
| **Duplicate words in the results of the word ‘القران’ ‘the Quran’** | |
| 'بالقرآن’: with the Quran | ‘bialqaran’ |
| 'قرآن': Quran | ‘qirana’ |
| 'للقرآن': for the Quran | ‘lilaquran’ |
| 'تلاوته': its recitation | ‘tilawth’ |

| | |
|---|---|
| 'تنزيله': its revelation | ‘tinzilh’ |
| 'قرآنا': Quran | ‘qirana’ |
| 'والقرآن': and the Quran | ‘walqaran’ |
| 'القران': the Quran | ‘alqran’ |
| 'كتابه': its book | ‘kitabh’ |

These drawbacks occurred because the raw text was tokenised using a whitespace tokenizer. The whitespace tokenizer splits text into words whenever it finds a whitespace character. The training dataset of raw Arabic text needs further morphological analysis to overcome duplication in the result and thus enhance the accuracy of the model. For example, the tokenisation of the word ‘بالقران’ based on POS tagging is [(‘ب’, ‘IN’), (‘ال’, ‘DT’), (‘قران’, ‘NNP’)]. This semantic similar word model might be used in the future to develop a new Arabic WordNet for the Islamic domain.

## 7.5 Challenges

The challenges I faced during the conduct of this research are as follows:

The lack of Quran question–answer datasets negatively affects the evaluation accuracy and credibility of Quranic search tools. This issue makes the assessment of different search methods for the Quran inaccurate and imprecise. The majority of previous studies used question–answer datasets which are different in terms of size, question types, answer types and type of retrieved knowledge, affecting the accuracy of the evaluation results. According to the literature, there is no gold standard dataset for Arabic questions and answers, particularly in the domain of the Quran, which can be used by researchers to evaluate their proposed search methods. For this reason, I developed the Arabic annotated corpus for Quranic questions and answers to serve as the benchmark for testing any Quranic search tools.

The Quran script (Othmani text) has many different words from standard Quranic Arabic script. Additionally, current modern standard Arabic MSA language has some differences from the classical Arabic language of the Quran. This issue was discussed using a classic Arabic language script for the Quran and the description of each verse in modern standard Arabic language called *tafseer*, such as *tafseer almuassar* and *tafseer aljalalyan.*

Free Arabic natural language processing tools are limited and still have some drawbacks, such as their stem and the lemma of the Arabic token. Examples of these tools are a part of speech taggers, the Arabic WordNet, especially in the domain of the Quranic sciences, and the Arabic morphological analyser (stemmer and lemmatisation). This affects the result accuracy of the proposed Arabic Quranic semantic search model.

Because of the lack of an Arabic Quranic ontology, current Arabic Quran ontologies have different scopes, formats, entity names and text languages. Additionally, a single Quranic ontology does not cover most of the knowledge in the Quran. Therefore, these ontologies need to be increased, normalised and combined with other Quran resources, such as Quran word meanings. The new Quran ontology is constructed from more than 10 datasets related to the Quran. This ontology has more than 1,070,000 triples, 78 classes, 51 object properties, 34 data properties and 86,588 individuals.

## 7.6 Conclusion

This thesis has achieved its aims and objectives which were stated in the first chapter. It evaluated existing search tools constructed for the Holy Quran against 13 criteria in terms of search features, output features, precision of the retrieved verses, recall database size and types of database content.

Then, the study reviewed existing Quran ontologies and compared them against 11 criteria. Some deficits have been found in all these ontologies. Additionally, a single Quranic ontology does not cover most of the knowledge in the Quran. Therefore, I developed a new Arabic–English Quran ontology from 10 datasets related to the Quran, such as the Quran chapter and verse names, Quran word meanings and Quran topics. The main aim of developing a Quranic ontology is to facilitate the retrieval of knowledge from the Quran. Additionally, the Quran ontology will enrich raw Arabic and English Quran text with Islamic semantic tags.

Furthermore, this research developed the first Annotated Corpus of Arabic Questions and Answers on the Quran. This corpus has 2,200 pairs of questions and answers collected from trusted Islamic sources. Each pair of question and answer is labelled with five tags. Examples of tags are question type (either factoid or descriptive), topic of question based on the Quran ontology and question class.

Finally, this thesis explained a new semantic search model for the Arabic Quran based on my Quran ontology. This model aims to address limitations in existing Quran search applications. The search tool uses both information retrieval techniques and semantic search technologies. The performance of this search model is evaluated using the Annotated Corpus of Arabic Questions and Answers on the Quran.

## List of References

’Ashur, Q. (2001). *1000 سؤال وجواب في القرآن الكريم*. Beirut: Dar Ibn Hazm.

(Indonesia), L. P. M. A.-Q., & Agama, I. D. (2010). *Syaamil al-Qur’an miracle the reference: mudah, sahih, lengkap, dan komprehensif*. Sygma Pub. Retrieved from https://books.google.co.uk/books?id=ciijnQAACAAJ

Abbas, N., Aldhubayi, L., Al-Khalifa, H., Alqassem, Z., Atwell, E. S., Dukes, K., … Sharaf, M. (2012). Unifying linguistic annotations and ontologies for the Arabic Quran. In *Proc WACL2 Second Workshop on Arabic Corpus Linguistics* (p. 13). Leeds.

Abbas, N. H. (2009a). *Quran’search for a Concept’Tool and Website*. *Unpublished MSc Dissertation, University of Leeds*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Quran+?Search+for+a+Concept?+Tool+and+Website#0

Abbas, N. H. (2009b). *Quran ‘search for a concept’ tool and website*. Retrieved from http://quranykeywords.appspot.com/

Abdelnasser, H., Mohamed, R., Ragab, M., Mohamed, A., Farouk, B., El-Makky, N., & Torki, M. (2014). Al-Bayan: An Arabic Question Answering System for the Holy Quran. In *EMNLP Workshop on Arabic Natural Language Processing (ANLP)* (p. 57).

Abdul-Rahman, M. S. (2009). *Tafsir Ibn Kathir Juz’ 1 (Part 1): Al-Fatihah 1 to Al-Baqarah 141 2nd Edition*. MSA Publication Limited. Retrieved from https://books.google.co.uk/books?id=sIryWTYotyYC

Abraham, B., Esther, K., & Christian, K. (2005). Querying the semantic web with ginseng: A guided input natural language search engine. In *5th Workshop on Information Technologies and Systems* (pp. 112–126). Las Vegas, NV. https://doi.org/10.1.1.90.7212

Abu Shawar, B., & Atwell, E. (2004). An Arabic chatbot giving answers from the Qur’an. *Proceedings of TALN04*, *2*, 197–202.

Ahmad, O., Hyder, I., Iqbal, R., Murad, M. A. A., Mustapha, A., Sharef, N. M., & Mansoor, M. (2013). A Survey of Searching and Information Extraction on a Classical Text Using Ontology-based semantics modeling: A Case of Quran. *Life Science Journal*, *10*(4).

Al-Fadili, A. H. (1980). *مختصر النحو* (seven). Jeddah: DAr Ashrooge.

Al-Yahya, M., Al-Khalifa, H., Bahanshal, A., Al-Odah, I., & Al-Helwah, N. (2010). An ontological model for representing semantic lexicons: an application on time nouns in the holy Quran. *Arabian Journal for Science and Engineering*, *35*(2), 21.

Al Chalabi, H. M., Ray, S. K., & Shaalan, K. (2015). Question classification for Arabic question answering systems. In *Information and Communication Technology Research (ICTRC)* (pp. 310–313). IEEE.

Al Gharaibeh, A., Al Taani, A., & Alsmadi, I. (2011). The Usage of Formal Methods in Quran Search System. In *Proceedings of International Conference on Information and Communication Systems, Ibrid, Jordan* (pp. 22–24).

Alfaifi, A. Y. G. (2015). *Building the Arabic Learner Corpus and a System for Arabic Error Annotation*. University of Leeds.

Allemang, D., & Hendler, J. (2008). *Semantic web for the working ontologist : modeling in RDF, RDFS and OWL*. *Journal of empirical research on human research ethics JERHRE* (Vol. 6). https://doi.org/10.1525/jer.2011.6.3.toc

Alobaid, A. A. (n.d.). *OnToology: An online tool for ontology documentation and evaluation*. Retrieved from http://oa.upm.es/41331/1/TFM_AHMAD_ADEL_ALOBAID.pdf

Alqahtani, M, & Atwell, E. (2018). Developing Bilingual Arabic-English Ontologies of Al-Quran. In *2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)* (pp. 96–101). London: IEEE.

Alqahtani, Mohammad, & Atwell, E. (2017). Evaluation Criteria for Computational Quran Search. *International Journal on Islamic Applications in Computer Science And Technology*, *5*(1). Retrieved from http://www.sign-ific-ance.co.uk/index.php/IJASAT/article/view/1562

Alrehaili, S. M., & Atwell, E. (2014). Computational ontologies for semantic tagging of the Quran: A survey of past approaches. *LREC 2014 Proceedings*.

Alsheddi, A. S. (2016). *Edit distance adapted to natural language words towards an unsupervised computation of the morphological relatedness : Arabic as case study*. Imam Muhammad ibn Saud Islamic University.

Amerland, D. (2013). *Google Semantic Search: Search Engine Optimization (SEO) Techniques That Get Your Company More Traffic, Increase Brand Impact, and Amplify Your Online Presence*. Que Publishing. Retrieved from https://books.google.com/books?id=niCJxqdc-eoC&pgis=1

Antoniou, G. (Grigoris), & Frank,  van H. (2012). *A Semantic Web primer* (second). London, England: The MIT Press.

Atwell, E., Brierley, C., Dukes, K., Sawalha, M., & Sharaf, A.-B. (2011). An Artificial Intelligence approach to Arabic and Islamic content on the internet. In *Proceedings of NITS 3rd National Information Technology Symposium*.

Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (2012). Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory*, *18*(3), 257–279. https://doi.org/10.1007/s10588-012-9121-2

Berger, A., Caruana, R., Cohn, D., Freitag, D., & Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 192–199). ACM.

Bernhard, D., & Gurevych, I. (2009). Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* (pp. 728–736). Association for Computational Linguistics.

Campo, J. E. (2009). *Encyclopedia of Islam*. Infobase Publishing.

Castells, P., Fernández, M., & Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *19*(2), 261–272. https://doi.org/10.1109/TKDE.2007.22

Cleverdon, C. W., Mills, J., & Keen, M. (1966). Factors determining the performance of indexing systems.

Damljanovic, D. (2012). FREyA: An interactive way of querying Linked Data using natural language. In *Extended Semantic Web Conference* (pp. 125–138). Springer Berlin Heidelberg.

Dar al-Maarifah. (1999). *Mushaf al Tajweed* (4th ed.). Damascus: Dar al-Maarifah. Retrieved from http://www.islamicbookstore.com/b8898.html

Dong, H., Hussain, F., & Chang, E. (2008). A survey in semantic search technologies. In *Digital Ecosystems and Technologies*. 2nd IEEE International Conference on. IEEE. Retrieved from http://espace.library.curtin.edu.au/R?func=dbin-jump-full&object_id=116029

Dukes, K. (2013). Statistical Parsing by Machine Learning from a Classical Arabic Treebank, *PhD Thesis*. Retrieved from http://corpus.quran.com/

Eclipse Deeplearning4j Development Team. (2017). Deeplearning4j. Retrieved 13 November 2017, from https://deeplearning4j.org

El-Naggar, Z. (2006). *Al-Hayawanat FI Al-Quran AL-KAREEM* (First). Beirut: DAR El-Marefah.

Euzenat, J., & Shvaiko, P. (2013). *Ontology matching Solutions to problems* (second). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38721-0

Fazzingaa, B., & Lukasiewiczb, T. (2010). Semantic search on the Web. *Semantic Web*, *1*(2), 89–96. https://doi.org/10.3233/SW-2010-0023

Fernández, M., Cantador, I., & López, V. (2011). Semantically enhanced Information Retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*452–434 ,(4)*9* ,.

Fox, E. A. (1983). *Characterization of two new experimental collections in computer* *and information science containing textual and bibliographic concepts*. Cornell University.

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI* (Vol. 7, pp. 1606–1611).

Garcia, A., O’Neill, K., Garcia, L. J., Lord, P., Stevens, R., Corcho, O., & Gibson, F. (2010). Developing ontologies within decentralised settings. In *Semantic e-Science* (pp. 99–139). Boston, MA.: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4419-5908-9_4

Ghidini, C., Kump, B., Lindstaedt, S., Mahbub, N., Pammer, V., Rospocher, M., & Serafini, L. (2009). MoKi: The enterprise modelling wiki. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5554 LNCS, pp. 831–835). Berlin: Springer. https://doi.org/10.1007/978-3-642-02121-3_65

Gómez-Pérez, A., & Suárez-Figueroa, M. C. (2008). NeOn methodology: scenarios for building networks of ontologies. *Poster and Demo*, 43.

Gruber, T. (2009). *Ontology in Encyclopedia of Database Systems*. (M. T. (Eds. . Liu, Ling, Özsu, Ed.), *Encyclopedia of Database Systems*. Springer Verlag. Retrieved from http://tomgruber.org/writing/ontology-definition-2007.htm

Habash, N., Rambow, O., & Roth, R. (2009). 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* (pp. 102–109).

Hakkoum, A., & Raghay, S. (2016). Semantic Q&amp;A System on the Qur’an. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-016-2251-y

Harmon, D. (1993). Overview of the First Text REtrieval Conference (TREC-1). *NIST Special Publication*, 207–500. Retrieved from https://trec.nist.gov/pubs/trec1/papers/01.txt

Hasan, A. M., & Zakaria, L. Q. (2016). QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE AND PATTERN MATCHING. *Journal of Theoretical & Applied Information Technology*, *87*(2).

Hitzler, P., Krotzsch, M., & Rudolph, S. (2009). *Foundations of semantic web technologies*. Chapman and Hall/CRC.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. Retrieved from http://arxiv.org/abs/1612.03651

Jurafsky, D., & Martin, J. H. (2017). Question Answering. In *Speech and Language Processing* (third, pp. 1–20). Retrieved from https://web.stanford.edu/~jurafsky/slp3/28.pdf

Kassel, G. (2005). *Integration of the DOLCE top-level ontology into the OntoSpec methodology*. *LaRIA RESEARCH REPORT*. Amiens. Retrieved from https://arxiv.org/abs/cs/0510050

Kaufmann, E., Bernstein, A., & Fischer, L. (2007). NLP-Reduce: A "naıve" but Domain-independent Natural Language Interface for Querying Ontologies. In *4th European Semantic Web Conference ESWC* (pp. 1–2).

Keet, C. M. (2012). Transforming semi-structured life science diagrams into meaningful domain ontologies with DiDOn. *Journal of Biomedical Informatics*, *45*(3), 482–494. https://doi.org/10.1016/j.jbi.2012.01.004

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Routledge. https://doi.org/10.4324/9781315843674

Khan, H. U., Saqlain, S. M., Shoaib, M., & Sher, M. (2013). Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*, *2*(6), 570–575.

Lebret, R. P. (2016). *Word Embeddings for Natural Language Processing*. Ecole Polytechnique Fédérale de Lausanne.

Lei, Y., Uren, V., & Motta, E. (2006). Semsearch: A search engine for the semantic web. In Steffen Staab & V. Svátek (Eds.), *Managing Knowledge in a World of Network*. Springer Berlin Heidelberg. https://doi.org/10.1007/11891451_22

Li, X., & Dan, R. (2002). Learning Question Classifiers. *COLING '02 Proceedings of the 19th International Conference on Computational Linguistics*, 1–7. https://doi.org/10.3115/1072228.1072378

Li, X., & Roth, D. (1998). Learning Question Classifiers: The Role of Semantic Information. *Natural Language Engineering*, *1*(1), 0–0. Retrieved from https://pdfs.semanticscholar.org/f2b1/8264de28827a061fe9e22c437d1f616fdb4a.pdf

Li, Y., & Yang, T. (2017). *Word Embedding for Understanding Natural Language: A Survey* (Vol. 26). https://doi.org/10.1007/978-3-319-53817-4

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The {Stanford} {CoreNLP} Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (pp. 55–60). Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word*

*Representations in Vector Space*. Retrieved from http://ronan.collobert.com/senna/

Mony, M., Rao, J., & Potey, M. (2014). Semantic Search based on Ontology Alignment for Information Retrieval. *International Journal of Computer Applications* , 10)*107*).

Muhammad, A. (2012). *Annotation of conceptual co-reference and text Mining the Qur'an*. Retrieved from http://etheses.whiterose.ac.uk/id/eprint/4160

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., & Randeree, B. (2015). Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 269–281). Retrieved from http://www.aclweb.org/anthology/S15-2047

Obrst, L., Ashpole, B., Ceusters, W., Mani, I., & Smith, B. (2007). The Evaluation of Ontologies: Toward Improved Semantic Interoperability. In *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences* (pp. 1–19). New York: Springer Verlag. https://doi.org/10.1007/978-0-387-48438-9_8

Ou, S., Pekar, V., Orasan, C., Spurk, C., & Negri, M. (2008). Development and Alignment of a Domain-Specific Ontology for Question Answering.. In *LREC*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.485.3339&rep=rep1&type=pdf

Paşca, M. (2003). Open-Domain Question Answering from Large Text Collections. *Computational Linguistics*, *29*(4), 665–667. https://doi.org/10.1162/089120103322753383

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. Retrieved from http://arxiv.org/abs/1806.03822

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Retrieved from https://arxiv.org/pdf/1606.05250.pdf

Raza, S. A., Rehan, M., Farooq, A., Ahsan, S. M., & Khan, M. S. (2014). AN ESSENTIAL FRAMEWORK FOR CONCEPT BASED EVOLUTIONARY QURANIC SEARCH ENGINE (CEQSE).

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems*101–96 ,(3)*21* ,. https://doi.org/10.1109/MIS.2006.62

Sharaf, A.-B. M., & Atwell, E. (2012). QurAna: Corpus of the Quran annotated with Pronominal Anaphora. In *LREC* (pp. 130–137). Citeseer.

Sharaf, A.-B. M., & Atwell, E. S. (2006). QurSim: A corpus for evaluation of relatedness in short texts. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/190_Paper.pdf

Sherif, M. A., & Ngonga Ngomo, A.-C. (2009). Semantic QuranA multilingual resource for natural-language processing. *Semantic Web*.

Shoaib, M., Nadeem Yasin, M., Hikmat, U. K., Saeed, M. I., & Khiyal, M. S. H. (2009). Relational WordNet model for semantic search in Holy Quran. In *Emerging Technologies, 2009. ICET 2009. International Conference on* (pp. 29–34). IEEE.

Singhal, A. (2001). Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *24*, 4.

Skuce, D. (1995). Conventions for reaching agreement on shared ontologies. In *Proceedings of the 9th Knowledge Acquisition for Knowledge Based Systems Workshop*.

Spiteri, L., & Richard, N. (2013). Evaluation of Internet Search Engines: Methodological Issues and Assumptions. In *the Annual Conference of CAIS*. Retrieved from http://www.cais-acsi.ca/ojs/index.php/cais/article/view/486

Staab, S, & Studer, R. (2001). Knowledge processes and ontologies. *IEEE Intelligent Systems*, *16*(1), 26–34. Retrieved from https://ieeexplore.ieee.org/iel5/5254/19693/00912382.pdf

Stragand, G. (2011). Simple Fuzzy String Similarity in Java - CodeProject. Retrieved 2 October 2017, from https://www.codeproject.com/Articles/147230/Simple-Fuzzy-String-Similarity-in-Java

Sudeepthi, G., Anuradha, G., & Babu, M. S. P. (2012). A Survey on Semantic Web Search Engine. *International Journal of Computer Science*, *9*.

Ta’a, A., Abdullah, M. S., Ali, A. B. M., & Ahmad, M. (2014). Themes-based classification for Al-Quran knowledge ontology. *International Conference on ICT Convergence*, (074), 89–94. https://doi.org/10.1109/ICTC.2014.6983090

Vrandeči, D. (2010). *Ontology Evaluation*. The Karlsruhe Institute of Technology.

W3C, T. W. W. W. C. (2015). Ontologies. Retrieved from http://www.w3.org/standards/semanticweb/ontology

Waheeb, A., & Babu, A. P. (2016). Classification of arabic questions using multinomial naive bayes and support vector machines. *INTERNATIONAL JOURNAL OF*

*LATEST TRENDS IN ENGINEERING AND TECHNOLOGY*, (SACAIM), 82–86. https://doi.org/10.21172

Yahya, Z., Abdullah, M. T., Azman, A., & Kadir, R. A. (2013). Query Translation Using Concepts Similarity Based on Quran Ontology for Cross-Language Information Retrieval. *Journal of Computer Science*, *9*(7), 889–897. https://doi.org/10.3844/jcssp.2013.889.897

Yauri, A., Azman, A., Murad, M. A. A., & Kadir, A. (2014). Semantic Web Application for Historical Concepts Search in Al-Quran. *International Journal on Islamic Applications in Computer Science And Technology*, *2*(2), 1–7.

Yauri, A. R. (2014). *Automated Semantic Query Formulation for Quranic Verse Retrieval*. *Computer Science and IT*. Putra Malaysia, Malaysia.

Yauri, A. R., Kadir, R. A., Azman, A., & Murad, M. A. A. (2013). Quranic Verse Extraction base on Concepts using OWL-DL Ontology. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(23), 4492–4498.

Yu, L. (2014). *A Developer's Guide to the Semantic Web*. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43796-4

Yunus, M. A., Zainuddin, R., & Abdullah, N. (2010). Semantic query for Quran documents results. In *Open Systems (ICOS), 2010 IEEE Conference on* (pp. 1–5). IEEE.

Zaeri, A., & Nematbakhsh, M. (2015). A Semantic Search Algorithm for Ontology Matching. *Semantic Web Journal Net254* ,.

Zhu, M. (2004). Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, *2*, 30.

# Appendix A
# Evaluation of Quran Search Tools

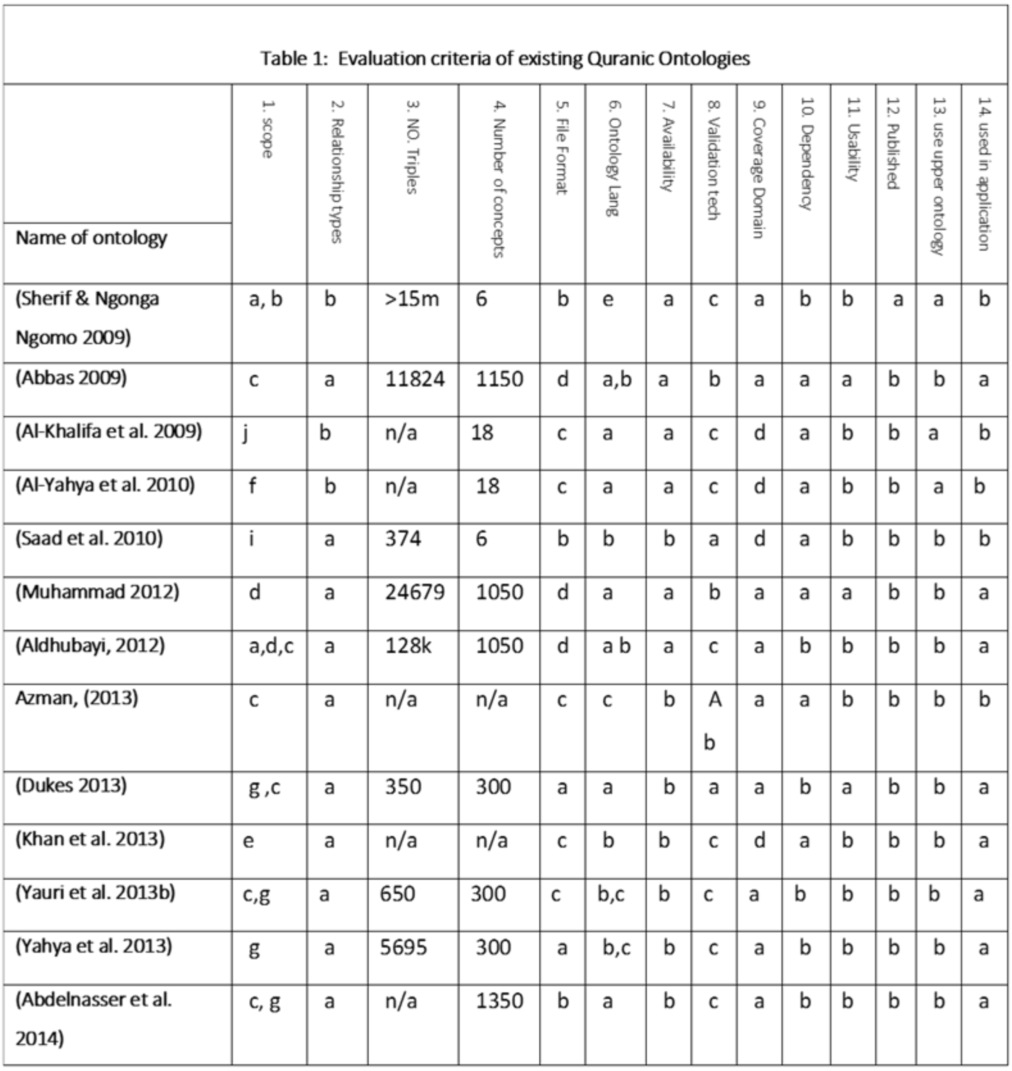
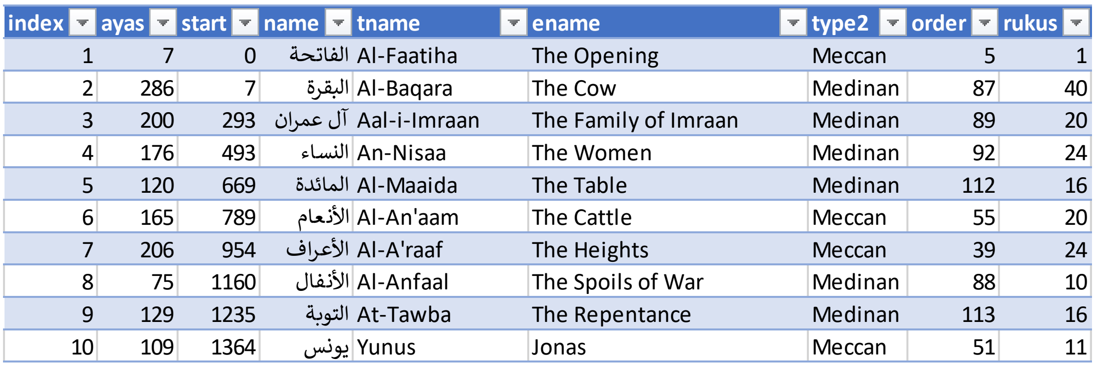


Table 1: Evaluation criteria of existing Quranic Ontologies

| Name of ontology | 1. scope | 2. Relationship types | 3. NO. Triples | 4. Number of concepts | 5. File Format | 6. Ontology Lang | 7. Availability | 8. Validation tech | 9. Coverage Domain | 10. Dependency | 11. Usability | 12. Published | 13. use upper ontology | 14. used in application |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (Sherif & Ngonga Ngomo 2009) | a, b | b | >15m | 6 | b | e | a | c | a | b | b | a | a | b |
| (Abbas 2009) | c | a | 11824 | 1150 | d | a,b | a | b | a | a | a | b | b | a |
| (Al-Khalifa et al. 2009) | j | b | n/a | 18 | c | a | a | c | d | a | b | b | a | b |
| (Al-Yahya et al. 2010) | f | b | n/a | 18 | c | a | a | c | d | a | b | b | a | b |
| (Saad et al. 2010) | i | a | 374 | 6 | b | b | b | a | d | a | b | b | b | b |
| (Muhammad 2012) | d | a | 24679 | 1050 | d | a | a | b | a | a | a | b | b | a |
| (Aldhubayi, 2012) | a,d,c | a | 128k | 1050 | d | a b | a | c | a | b | b | b | b | a |
| Azman, (2013) | c | a | n/a | n/a | c | c | b | A<br>b | a | a | b | b | b | b |
| (Dukes 2013) | g ,c | a | 350 | 300 | a | a | b | a | a | b | a | b | b | a |
| (Khan et al. 2013) | e | a | n/a | n/a | c | b | b | c | d | a | b | b | b | a |
| (Yauri et al. 2013b) | c,g | a | 650 | 300 | c | b,c | b | c | a | b | b | b | b | a |
| (Yahya et al. 2013) | g | a | 5695 | 300 | a | b,c | b | c | a | b | b | b | b | a |
| (Abdelnasser et al. 2014) | c, g | a | n/a | 1350 | b | a | b | c | a | b | b | b | b | a |

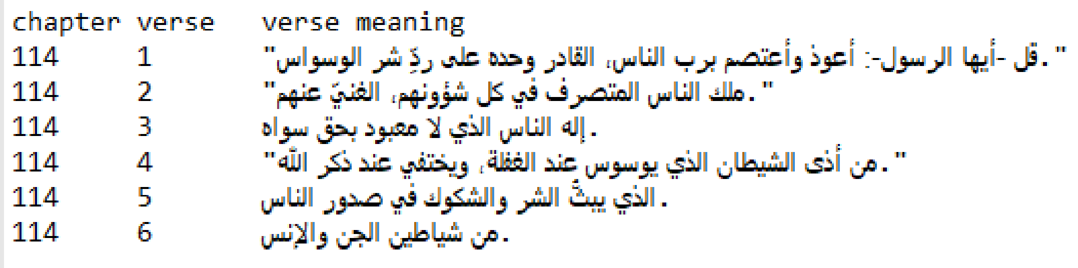
# Appendix B
# Examples of  the Quran Datasets used to Develop Ontology

## B.1  Quran Structure dataset



| index | ayas | start | name | tname | ename | type2 | order | rukus |
|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 0 | الفاتحة | Al-Faatiha | The Opening | Meccan | 5 | 1 |
| 2 | 286 | 7 | البقرة | Al-Baqara | The Cow | Medinan | 87 | 40 |
| 3 | 200 | 293 | آل عمران | Aal-i-Imraan | The Family of Imraan | Medinan | 89 | 20 |
| 4 | 176 | 493 | النساء | An-Nisaa | The Women | Medinan | 92 | 24 |
| 5 | 120 | 669 | المائدة | Al-Maaida | The Table | Medinan | 112 | 16 |
| 6 | 165 | 789 | الأنعام | Al-An'aam | The Cattle | Meccan | 55 | 20 |
| 7 | 206 | 954 | الأعراف | Al-A'raaf | The Heights | Meccan | 39 | 24 |
| 8 | 75 | 1160 | الأنفال | Al-Anfaal | The Spoils of War | Medinan | 88 | 10 |
| 9 | 129 | 1235 | التوبة | At-Tawba | The Repentance | Medinan | 113 | 16 |
| 10 | 109 | 1364 | يونس | Yunus | Jonas | Meccan | 51 | 11 |

## B.2 Quran description by Tafsir al-Jalalyen dataset



```
chapter verse    verse meaning
114     1        "قل -أيها الرسول-: أعوذ وأعتصم برب الناس، القادر وحده على ردِّ شر الوسواس."
114     2        "ملك الناس المتصرف في كل شؤونهم، الغنيِّ عنهم."
114     3        إله الناس الذي لا معبود بحق سواه.
114     4        "من أذى الشيطان الذي يوسوس عند الغفلة، ويختفي عند ذكر الله."
114     5        الذي يبثُّ الشر والشكوك في صدور الناس.
114     6        من شياطين الجن والإنس.
```

## B.3 Quran parallel words Translation dataset

| Arabic Text | Lemma | Othmani Text | English text Of the Quran | Transliteration | Location (chapter verse word) |
|---|---|---|---|---|---|
| وحصل | حصل | وَحُصِّلَ | And is made apparent | waḥuṣṣila | 100-10-1 |
| ما | ما | مَا | what | mā | 100-10-2 |
| في | فى | فِى | (is) in | fī | 100-10-3 |
| الصدور | صدر | ٱلصُّدُورِ | the breasts? | l-ṣudūri | 100-10-4 |
| والعاديات | عديت | وَٱلْعَٰدِيَٰتِ | By the racers | wal-ʿādiyāti | 100-1-1 |
| إن | ان | إِنَّ | Indeed | inna | 100-11-1 |
| ربهم | رب | رَبَّهُم | their Lord | rabbahum | 100-11-2 |
| بهم | بهم | بِهِمْ | about them | bihim | 100-11-3 |
| يومئذ | يوميذ | يَوْمَئِذٍۢ | that Day | yawma-idhin | 100-11-4 |
| لخبير | خبير | لَّخَبِيرٌۢ | (is) surely All-Aware | lakhabīrun | 100-11-5 |
| ضبحا | ضبح | ضَبْحًۭا | panting | ḍabḥan | 100-1-2 |
| فالموريات | موريت | فَٱلْمُورِيَٰتِ | And the producers of sparks | fal-mūriyāti | 100-2-1 |
| قدحا | قدح | قَدْحًۭا | striking | qadḥan | 100-2-2 |
| فالمغيرات | مغيرت | فَٱلْمُغِيرَٰتِ | And the chargers | fal-mughīrāti | 100-3-1 |
| صبحا | صبح | صُبْحًۭا | (at) dawn | ṣub'ḥan | 100-3-2 |
| فأثرن | اثار | فَأَثَرْنَ | Then raise | fa-atharna | 100-4-1 |
| به | به | بِهِۦ | thereby | bihi | 100-4-2 |
| نقعا | نقع | نَقْعًۭا | dust | naqʿan | 100-4-3 |
| فوسطن | وسط | فَوَسَطْنَ | Then penetrate (in the) centre | fawasaṭna | 100-5-1 |
| به | به | بِهِۦ | thereby | bihi | 100-5-2 |
| جمعا | جمع | جَمْعًا | collectively | jamʿan | 100-5-3 |
| إن | ان | إِنَّ | Indeed | inna | 100-6-1 |
| الإنسان | انسن | ٱلْإِنسَٰنَ | mankind | l-insāna | 100-6-2 |
| لربه | رب | لِرَبِّهِۦ | to his Lord | lirabbihi | 100-6-3 |
| لكنود | كنود | لَكَنُودٌۭ | (is) surely ungrateful | lakanūdun | 100-6-4 |
| وإنه | ان | وَإِنَّهُۥ | And indeed he | wa-innahu | 100-7-1 |
| على | على | عَلَىٰ | on | ʿalā | 100-7-2 |
| ذلك | ذلك | ذَٰلِكَ | that | dhālika | 100-7-3 |
| لشهيد | شهيد | لَشَهِيدٌۭ | surely (is) a witness | lashahīdun | 100-7-4 |
| وإنه | ان | وَإِنَّهُۥ | And indeed he | wa-innahu | 100-8-1 |
| لحب | حب | لِحُبِّ | in (the) love | liḥubbi | 100-8-2 |
| الخير | خير | ٱلْخَيْرِ | (of) wealth | l-khayri | 100-8-3 |

## B.4 Examples from datasets of Quran Arabic word Meaning dataset

| Verse | chapter | Quran Word | word meaning |
|---|---|---|---|
| 2 | 112 | الله الصّمد | هو وَحْدَه المقصود في الحَوائج |
| 4 | 112 | كُفُوًا | مُـكافِـئـًـا ومُـمَـاثِلاً |
| 1 | 113 | أعوذ | أعْتـَصِمُ وأسْتجير |
| 1 | 113 | بربّ الفلق | بربّ الصّبْح. أو الخَـلـْق كلـّهمْ |
| 3 | 113 | شرّ غاسق | شرّ الليل |
| 3 | 113 | وَقب | دَخَل ظلامه في كلّ شيء |
| 4 | 113 | النفـاثـات في العقد | النِّساء السَّـواحِر يَـنـْـفُـثـْنَ في عُـقـَـد الخيْط حين يَسْحَرْنَ |
| 1 | 114 | أعوذ | أعْـتصِـمُ وأستجير |
| 1 | 114 | بربّ النـّاس | مُرَبِّـيهِمْ ومُدبّر أحوالهم |
| 2 | 114 | مَـلِـكِ النّـاس | مالِكِهِمْ مِلـْكًـا تـامّـا |
| 3 | 114 | إلـه النّاس | مَعْبُودِهِم الحقّ |
| 4 | 114 | الوَسواس | المُوَسْوس جـِنـِّيّـا أو إنـْسِـيّـا |
| 4 | 114 | الخنّاس | المُـتـوَاري المُـخْـتـفي |
| 6 | 114 | الجـِنّة | الجـِنّ |

## B.5 Examples from datasets of Quran chapters' other names

| اسم السورة | **surah_Names** | | |
|---|---|---|---|
| الفاتحة | فاتحة الكتاب | fatha alkataab | fatha alkatab |
| الفاتحة | فاتحة القران | fatha alqaraan | fatha alqaran |
| الفاتحة | أم الكتاب | 'am alkataab | 'am alkatab |

| | | | |
|---|---|---|---|
| الفاتحة | أم القران | 'am alqaraan | 'am alqaran |
| الفاتحة | القران العظيم | alqaraan al'adheem | alqaran aladhim |
| الفاتحة | السبع المثاني | assab' almathaanee | assab almathani |
| الفاتحة | الحمد | alhamd | alhamd |
| البقرة | الزهراوين | azzahraaween | azzahrawin |
| البقرة | النساء الطولى | annasaa' attawlaa | annasa' attawla |
| البقرة | سنام القران | sanaam alqaraan | sanam alqaran |
| ال عمران | الزهراوين | azzahraaween | azzahrawin |
| ال عمران | الكنز | alkanz | alkanz |
| ال عمران | المعينة | alma'eena | almaina |
| ال عمران | المجادلة | almajaadla | almajadla |
| ال عمران | الإستغفار | al'istghfaar | al'istghfar |
| ال عمران | الأمان | al'amaan | al'aman |
| ال عمران | طيبة | tayba | tayba |
| النساء | النساء الكبرى | annasaa' alkabraa | annasa' alkabra |
| المائدة | المنقذة | almanqzha | almanqzha |
| المائدة | الأخيار | al'akhyaar | al'akhyar |
| المائدة | العقود | al'aqoud | alaqud |
| الأنفال | بدر | badr | badr |
| التوبة | براءة | baraa'a | bara'a |
| التوبة | الفاضحة | alfaadha | alfadha |
| التوبة | المقشقشة | almaqshqsha | almaqshqsha |
| التوبة | العذاب | al'azhaab | alazhab |
| التوبة | المنقرة | almanqra | almanqra |

## B.6 Examples from datasets of Quran verses (Ayat) names dataset

| اسم الاية | En_translitration | اسم السورة | من رقم الاية |
|---|---|---|---|
| ادم | adm | الأعراف | 189 |
| الأخوة | al'akhwa | الحجرات | 9 |
| الإذن | al'izhn | النور | 27 |
| الإذن | al'izhn | النور | 58 |
| الإذن بالقتال | al'izhn baalqtaal | الحجرات | 39 |
| الإذن في خروج النساء | al'izhn fee kharouj annasaa' | الأحزاب | 53 |
| الأذى | al'azhaa | البقرة | 222 |
| الحبس | alhabs | النساء | 15 |
| الإرتداد | al'irtdaad | المائدة | 54 |
| الردة | arrada | المائدة | 54 |
| الإستئذان | al'isti'zhaan | النور | 27 |
| الإستثناء | al'istthnaa' | هود | 107 |
| الإسترجاع | al'istrjaa' | البقرة | 155 |
| الصبر | assabr | البقرة | 155 |
| الإستغفار | al'istghfaar | التوبة | 113 |
| النهي عن الإستغفار | annahee 'an al'istghfaar | التوبة | 113 |
| الإستهزاء | al'isthzaa' | البقرة | 15 |
| الإستواء | al'istwaa' | الأعراف | 54 |
| الإستواء | al'istwaa' | يونس | 3 |
| الإستواء | al'istwaa' | الرعد | 2 |
| الإستواء | al'istwaa' | طه | 5 |
| الإستواء | al'istwaa' | الفرقان | 59 |
| الإستواء | al'istwaa' | السجدة | 4 |
| الإستواء | al'istwaa' | الحديد | 4 |
| الإسراء | al'israa' | الإسراء | 1 |
| الأسرى | al'asraa | الأنفال | 67 |
| الإسلام | al'islaam | ال عمران | 64 |
| الإسلام | al'islaam | ال عمران | 19 |

# Appendix C
# Quran Ontology examples

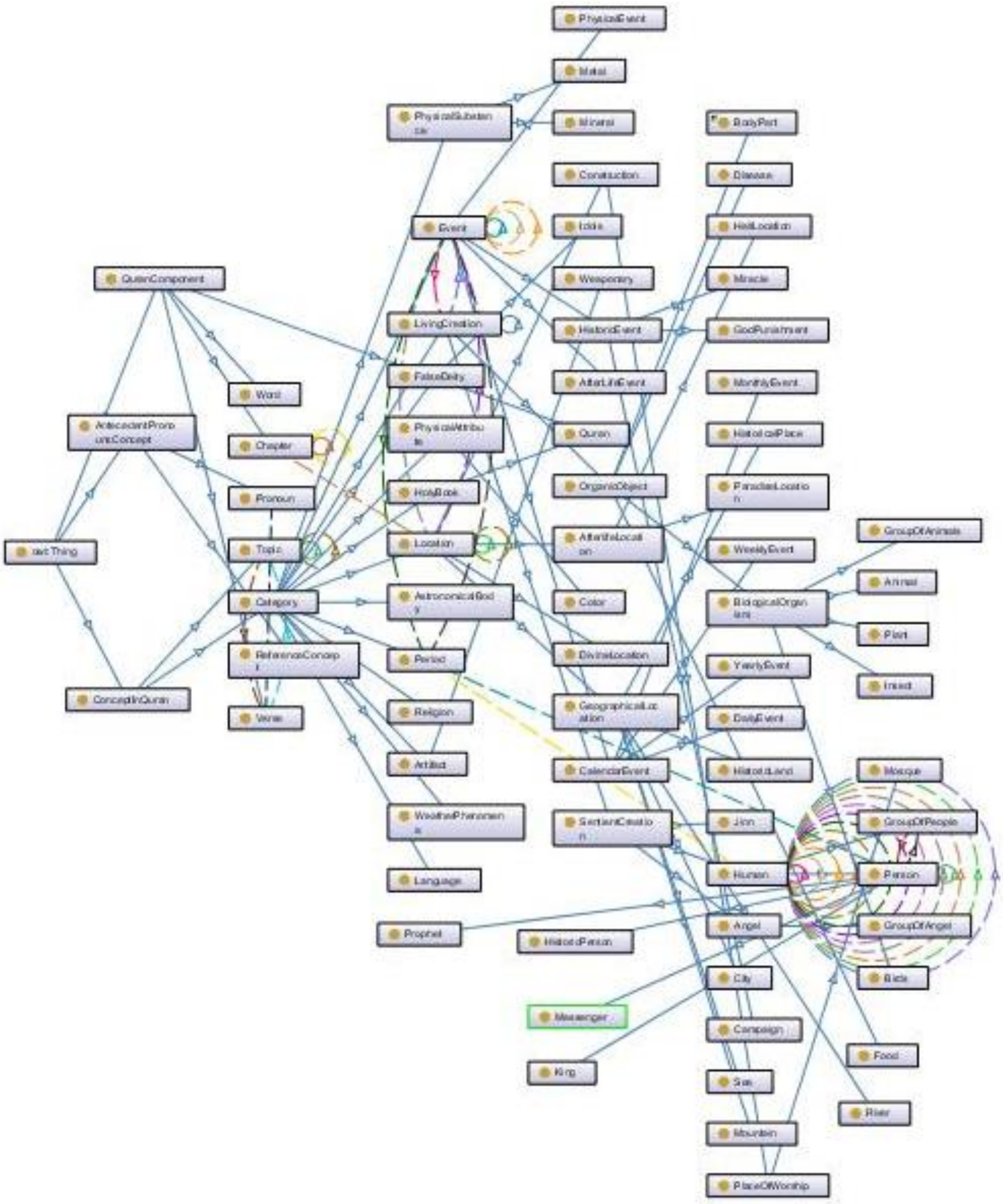
## C.1 Quran Ontology text file examples

All instances of class Weaponry: Arrow, Ladder,

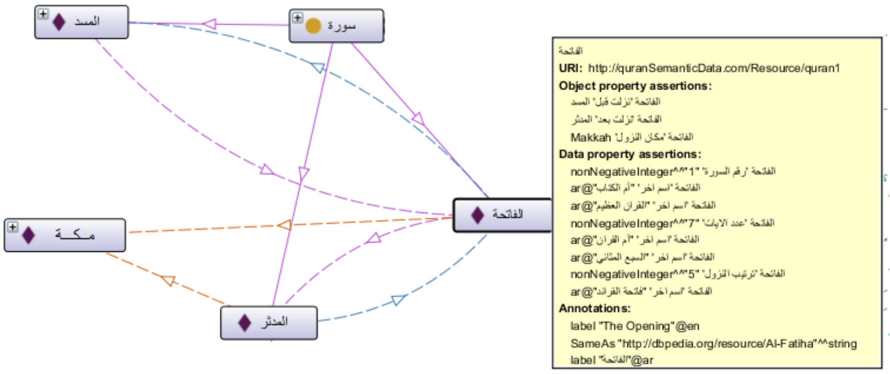
```
 <!-- http://QuranSemanticData.com/Resource/Artifact -->

  <owl:Class rdf:about="&Resource;Artifact">
      <rdfs:label xml:lang="ar">أداة أثرية</rdfs:label>
      <rdfs:label xml:lang="en">Artifact</rdfs:label>
      <rdfs:subClassOf rdf:resource="&Resource;Category"/>
      <owl:SameAs xml:lang="en">http://corpus.quran.com/concept.jsp?id=artifact</owl:SameAs>
  </owl:Class>

  <!-- http://QuranSemanticData.com/Resource/Weaponary -->

  <owl:Class rdf:about="&Resource;Weaponary">
      <rdfs:label xml:lang="ar">سلاح</rdfs:label>
      <rdfs:label xml:lang="en">Weaponry</rdfs:label>
      <rdfs:subClassOf rdf:resource="&Resource;Artifact"/>
      <owl:SameAs>http://corpus.quran.com/concept.jsp?id=weaponary</owl:SameAs>
      <owl:SameAs xml:lang="en">http://corpus.quran.com/concept.jsp?id=weaponary</owl:SameAs>
  </owl:Class>


<!-- http://QuranSemanticData.com/Resource/Knife -->

  <owl:NamedIndividual rdf:about="&Resource;Knife">
      <rdf:type rdf:resource="&Resource;Weaponary"/>
      <rdfs:label xml:lang="ar">سكينة</rdfs:label>
      <rdfs:label xml:lang="en">Knife</rdfs:label>
      <owl:SameAs>http://corpus.quran.com/concept.jsp?id=knife</owl:SameAs>
      <MentionedIn rdf:resource="&Resource;quran12-31"/>
  </owl:NamedIndividual>

  <!-- http://QuranSemanticData.com/Resource/Ladder -->

  <owl:NamedIndividual rdf:about="&Resource;Ladder">
      <rdf:type rdf:resource="&Resource;Artifact"/>
      <rdfs:label xml:lang="ar">سلم</rdfs:label>
      <rdfs:label xml:lang="en">Ladder</rdfs:label>
      <owl:SameAs>http://corpus.quran.com/concept.jsp?id=ladder</owl:SameAs>
      <MentionedIn rdf:resource="&Resource;quran6-35"/>
  </owl:NamedIndividual>

```

## C.2 The Quran Ontology graph

English Quran Ontology classes with their relationships

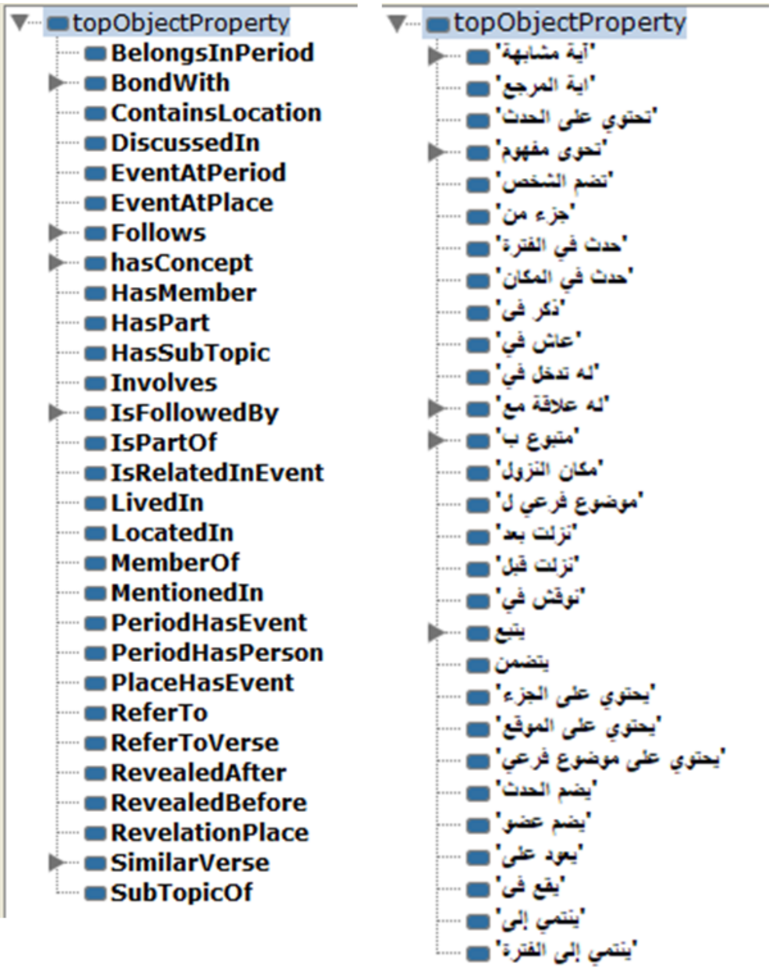


The Quranic chapter one ‘the opening’ graph



الفاتحة
**URI:** http://quranSemanticData.com/Resource/quran1
**Object property assertions:**
الفاتحة "نزلت قبل" المسد
الفاتحة "نزلت بعد" المدثر
الفاتحة "مكان النزول" Makkah
**Data property assertions:**
الفاتحة "رقم السورة" "1"^^nonNegativeInteger
الفاتحة "اسم اخر" "أم الكتاب"@ar
الفاتحة "اسم اخر" "القران العظيم"@ar
الفاتحة "عدد الايات" "7"^^nonNegativeInteger
الفاتحة "اسم اخر" "أم القران"@ar
الفاتحة "اسم اخر" "السبع المثاني"@ar
الفاتحة "ترتيب النزول" "5"^^nonNegativeInteger
الفاتحة "اسم اخر" "فاتحة القرائد"@ar
**Annotations:**
label "The Opening"@en
SameAs "http://dbpedia.org/resource/Al-Fatiha"^^string
label "الفاتحة"@ar

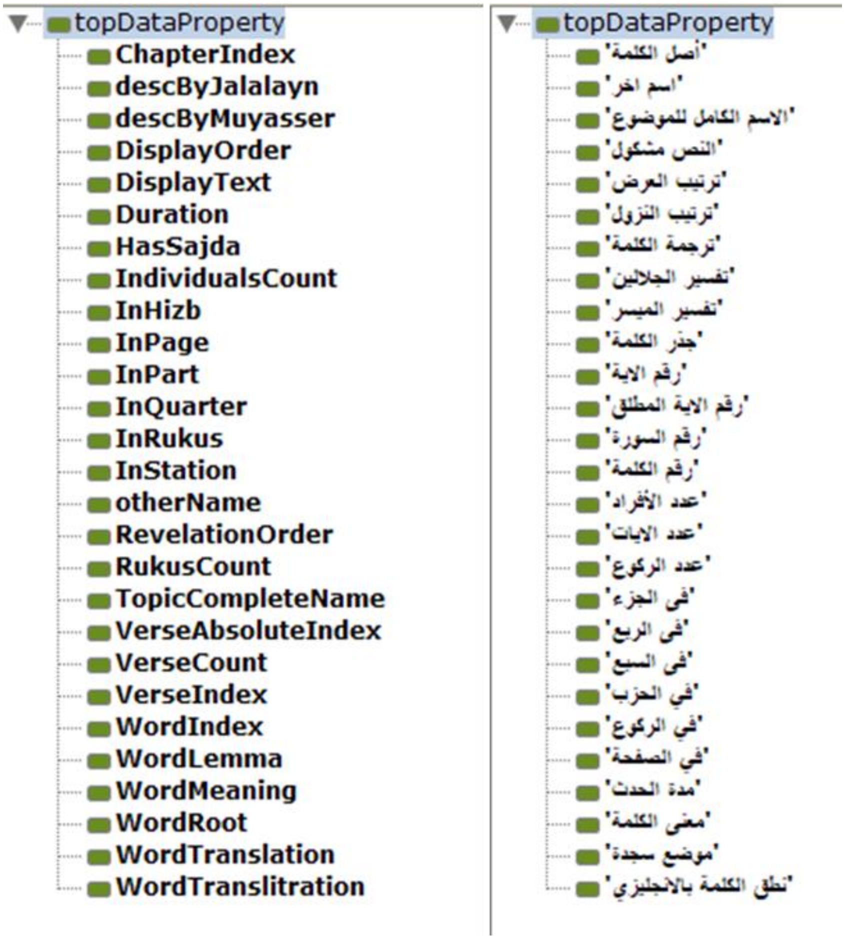
List of object properties in English and Arabic languages.



topObjectProperty
- BelongsInPeriod
- BondWith
- ContainsLocation
- DiscussedIn
- EventAtPeriod
- EventAtPlace
- Follows
- hasConcept
- HasMember
- HasPart
- HasSubTopic
- Involves
- IsFollowedBy
- IsPartOf
- IsRelatedInEvent
- LivedIn
- LocatedIn
- MemberOf
- MentionedIn
- PeriodHasEvent
- PeriodHasPerson
- PlaceHasEvent
- ReferTo
- ReferToVerse
- RevealedAfter
- RevealedBefore
- RevelationPlace
- SimilarVerse
- SubTopicOf

topObjectProperty
- 'آية مشابهة'
- 'اية المرجع'
- 'تحتوي على الحدث'
- 'تحوى مفهوم'
- 'تضم الشخص'
- 'جزء من'
- 'حدث في الفترة'
- 'حدث في المكان'
- 'ذكر فى'
- 'عاش في'
- 'له تدخل في'
- 'له علاقة مع'
- 'متبوع ب'
- 'مكان النزول'
- 'موضوع فرعي ل'
- 'نزلت بعد'
- 'نزلت قبل'
- 'نوقش في'
- يتبع
- يتضمن
- 'يحتوي على الجزء'
- 'يحتوي على الموقع'
- 'يحتوي على موضوع فرعي'
- 'يضم الحدث'
- 'يضم عضو'
- 'يعود على'
- 'يقع فى'
- 'ينتمي إلى'
- 'ينتمي إلى الفترة'

List of data properties in English and Arabic languages



- topDataProperty
  - ChapterIndex
  - descByJalalayn
  - descByMuyasser
  - DisplayOrder
  - DisplayText
  - Duration
  - HasSajda
  - IndividualsCount
  - InHizb
  - InPage
  - InPart
  - InQuarter
  - InRukus
  - InStation
  - otherName
  - RevelationOrder
  - RukusCount
  - TopicCompleteName
  - VerseAbsoluteIndex
  - VerseCount
  - VerseIndex
  - WordIndex
  - WordLemma
  - WordMeaning
  - WordRoot
  - WordTranslation
  - WordTranslitration

- topDataProperty
  - 'أصل الكلمة'
  - 'اسم اخر'
  - 'الاسم الكامل للموضوع'
  - 'النص مشكول'
  - 'ترتيب العرض'
  - 'ترتيب النزول'
  - 'ترجمة الكلمة'
  - 'تفسير الجلالين'
  - 'تفسير الميسر'
  - 'جذر الكلمة'
  - 'رقم الاية'
  - 'رقم الاية المطلق'
  - 'رقم السورة'
  - 'رقم الكلمة'
  - 'عدد الأفراد'
  - 'عدد الايات'
  - 'عدد الركوع'
  - 'فى الجزء'
  - 'فى الربع'
  - 'فى السبع'
  - 'في الحزب'
  - 'في الركوع'
  - 'في الصفحة'
  - 'مدة الحدث'
  - 'معنى الكلمة'
  - 'موضع سجدة'
  - 'نطق الكلمة بالانجليزي'

# Appendix D

## D.1 Web scraping questions and answers code from a website

```
# -*- coding: utf-8 -*-

from lxml import html
import requests
import sys

j=0

for i in range(1, 20):
    # copy of page content that has questions links
    page = requests.get('https://islamqa.info/ar/cat/242?page=' +str(i))
    tree = html.fromstring(page.content)

''''This will create a list of questions url after the main website url;
 html <a> with css class name:list-group-item and atribute herf
 '''
    links = tree.xpath('//a[@class="list-group-item"]/@href')

    #visit every question link per page
    for link in links:

        x ='https://islamqa.info' + link
        x= x.replace(" ", "")
        print repr(x)
        qpage=requests.get(x)
        # copy of question page content that has question and its answer
        qtree = html.fromstring(qpage.content)
        # extract question text and its answer
        questions = qtree.xpath('//div[@class="list-group-item"]//p/text()')
        #fn is file name i.e to save  each question-answer in a new file
        fn = 'quranQa/f'+str(j)+'.txt'
        j=j+1
        fo = open(fn, "wb")
        for qe in questions:
            qe =qe.encode('latin-1','ignore')
            fo.write(qe)

        fo.close()
```

## D.2 Extracting pairs of Question and Answer from a text file

```
# -*- coding: utf-8 -*-
"""
Created on Tue Nov 29 11:07:08 2016

@author: Mohammad
"""


import sys
import codecs
import re

a= open('QA_1000.txt', 'w',encoding="utf-8")


with codecs.open('1000QAinQuran.txt' ,'r',encoding='utf-8') as f:
    data = f.readlines()[0:]

    for line in data:
        line = line.replace('\n','')
        line = line.replace('\r','')
        s = line
        if (len(s) > 1):
            if(s.startswith('الباب') ):
                 a.write('\n' + line)
            elif (s.startswith('(س')):
                a.write('+' + line)
            elif (s.startswith('(ج')):
                a.write('+' + line)
            elif (re.match('(\((.+)\))', s)):
                a.write('\n' + line)

                #print ('\n' + line)
                 #a.write('\n' + line)
            else:
                #print ( line)
                a.write( line)

a.close()
```

## D.3 Extract Quran chapter and verses from questions in AQQAC

```python
''''' author:Mohammad Alqahtani
    extract Al Quran chapters and verses from each questions in AQQAC
    2017
'''

import csv
import sys
import codecs
import re

# save the result in new file C:\Quran_Q_A_Corpus\1000_q_a_V4.csv'
a= open('C:\\Quran_Q_A_Corpus\\1000_q_a_V4.csv', 'w',encoding="utf-8")

with codecs.open('C:\\Quran_Q_A_Corpus\\1000_ques_answer.csv' ,'r',encoding='utf-8') as f:
    data = f.readlines()[1:]
    raw=""
    c=0;
    for line in data:
        raw=""
        s = line.split(',')
        if (len(s) > 1):
            # extract chapter and verse e.g [4:الفاتحة]
            searchObj = re.findall( r'\[(.*?)\]', s[3] + s[4],)
            if (len(searchObj)>0):
             for x in searchObj:
             # add new line that contains question id and [chapter:verse]
                raw= raw +  s[0] + ';' + x +'\n'
             a.write(raw)
            else:
             raw=""+ s[0] +';' ""
             a.write( raw+'\n')
             print (raw)
             c= c+1;
        else:
             c= c+1;
    print(c)

a.close()
```

## D.4 Select best ontology concepts using cosine similarity

```python
# -*- coding: utf-8 -*-
"""
Created on Sun Feb  4 23:26:47 2018

@author: Mohammad
"""

import re, math
from nltk.stem.isri import ISRIStemmer
from collections import Counter
import codecs
WORD = re.compile(r'\w+')

def similarity_cosine(vec1, vec2):
     word_intersection = set(vec1.keys()) & set(vec2.keys())
     isMatch = (set(vec1.keys()) > set(vec2.keys())) or (set(vec1.keys()) < set(vec2.keys()))

    # compute consine similarity
     numerator_cosine = sum([vec1[x] * vec2[x] for x in word_intersection])

     sum_1 = sum([vec1[x]**2 for x in vec1.keys()])
     sum_2 = sum([vec2[x]**2 for x in vec2.keys()])

     denominator_cosine = math.sqrt(sum1) * math.sqrt(sum2)

     #similarity value is zero
     if not denominator_cosine:
        return (0.0,isMatch, word_intersection)
     else:
        similarity_value = float(numerator_cosine) / denominator_cosine

        return (similarity_value,isMatch, word_intersection)

def text_to_vector(text):
     words = WORD.findall(text)
     return Counter(words)

def stemWords(string1):
    xx = ISRIStemmer()
    txt1sp = string1.split(' ');
    newtext=""
    for t in txt1sp:
        newtext += xx.stem(t)+ ' '

    return(newtext)




a= open('C:\\Users\\Mohammad\\Google Drive\\resources\\Quran_Q_A_Corpus\\1000_q_a_OQC.csv', 'w',encoding="utf-8")

#with open('AllVersesNo.txt' ) as f:

```

```
with codecs.open('C:\\Users\\Mohammad\\Documents\\1000_q_a_OQC.txt' ,'r',encoding='utf-8') as f:
    data = f.readlines()[0:]
#print(data[0])
qid=0
cosSum=0
raw=""
raw1=""
for line in data:

        s = line.split(';"')
        if (len(s) > 1):
            if(len(s)>3):
                s[3]=s[3].replace('\r\n','')

            # vector1, vector2 using stem words of original text
            vector1 = text_to_vector(stemWords(s[3]))
            vector2 = text_to_vector((stemWords(s[1]+" "+s[2])))

            # vec1, vec2 using original text
            vec1 = text_to_vector((s[3]))
            vec2 = text_to_vector(((s[1]+s[2])))

            cosine = similarity_cosine(vector1, vector2)
            cosineStem = similarity_cosine(vec1, vec2)

            if (cosine[1] or cosineStem[1]):
                raw=""+s[0]+';'+s[3]+';'+ str(cosine[0])+';'+str(cosine[1])+';'
                      +', '.join(cosine[2])+';'+ str(cosineStem[0])+';'

                      + str(cosineStem[1])+';'+ ', '.join(cosineStem[2])
                )
                if (s[0] != qid):
                    a.write(raw1 +'\n')
                    cosSum = cosine[0]+cosineStem[0]
                    raw1=raw
                    qid=s[0]
                else:
                    if(cosSum< (cosine[0]+cosineStem[0])):
                       raw1=raw
a.write(raw1 +'\n')

a.close()

```

## D.5 Arabic Quranic Question classifier experiment code

```python
# -*- coding: utf-8 -*-
"""
Created on Fri Sep 28 02:07:06 2018

@author: Mohammad
"""


import numpy as np
import os
from random import shuffle
import re
import urllib.request
import zipfile
import lxml.etree
import codecs
from gensim.models import Word2Vec
import pickle
import fasttext


a = open('train.txt', 'w',encoding="utf-8")
b = open('test.txt', 'w',encoding="utf-8")

noise = re.compile(""" ّ    | # Tashdid
                             َ    | # Fatha
                             ً    | # Tanwin Fath
                             ُ    | # Damma
                             ؟     |
                             :    |
                             \d   |
                             \(   |
                             \)   |
                             \.   |
                             \n   |
                             \r   |
                             \!   |
                             {    |
                             }    |
                             ٌ    | # Tanwin Damm
                             ِ    | # Kasra
                             ٍ    | # Tanwin Kasr
                             ْ    | # Sukun
                             ـ     # Tatwil/Kashida
                         """, re.VERBOSE)

with codecs.open('C:\\Users\\Mohammad\\Google Drive\\resources\\QA_FT.txt' ,'r',encoding='utf-8') as f:
    data = f.readlines()
train_set=""
test_set =""
counter=1
for d in data:
    input_text = re.sub(r'["«,\.:\'\-!؟?\t؛،»}{\]\[] \)\(\d*', '', d)
    input_text = re.sub('\n', '', input_text)
    input_text = re.sub('\r', '', input_text)
    input_text = re.sub(noise, '', input_text)
    input_text = re.sub('س ', '', input_text)
    input_text = input_text.split(';')
    input_text[2] = re.sub(' ', '', input_text[2])
    if (counter % 10):
        print(input_text)
        #test_set.append(input_text[1] +' ' +input_text[2])
        a.write( input_text[2] +', ' +input_text[1] + '\n')
    else:
        b.write( input_text[2] +', ' +input_text[1] + '\n')
    counter=counter +1

a.close()
b.close()

classifier = fasttext.supervised('train.txt', 'model', label_prefix='__label__')

result = classifier.test('test.txt')
print ('P@1:', result.precision)
print ('R@1:', result.recall)
print ('Number of examples:', result.nexamples)

```

## D.6 Arabic Question particles list

| الاستفهام عن<br>Ask about | English translation | أسماء الاستفهام<br>Question particles | الحروف المقترنة بها<br>Connected Prepositions | عبارة الاستفهام<br>prepositional phrase | | | | |
|---|---|---|---|---|---|---|---|---|
| المفعولية<br>Entity | Whom<br>للعاقل | من<br>(men) | (bi)بـ<br>(Ann)عن<br>(fi)في<br>(li) لـ<br>(ila)إلى<br>With – about – in - for - to | بمن<br>With whom | عن من<br>About whom | فيمن<br>in whom | لمن<br>Whose | إلى من<br>to whom |
| | Which<br>عامة | أيّ<br>(ayu) | | بأيّ<br>with which | عن أيّ<br>About which | في أيّ<br>In which | لأيّ<br>For which | إلى أيّ<br>To which |
| | What<br>لغير العاقل | ما<br>(ma) | | بما<br>With what | عنما<br>About what | فيما<br>In what | لما<br>For what | إلى ما<br>To what |
| السببية<br>Reason (why) | What | ما<br>(ma) | لـ - بـ - في - من | لما<br>For what | بما<br>With what | فيما<br>In what | من ما<br>From what | |
| | which | أيّ<br>(ayu) | For – with – in - from | لأيّ<br>For which | بأيّ<br>with which | في أيّ<br>In which | من أيّ<br>from which | |
| | What/ why | ماذا<br>(matha) | | لماذا | بماذا | في ماذا | من ماذا | |
| الظرفية الزمانية أو المكانية<br>Adverb of Location and time | What | ما<br>(ma) | في - على - بـ | فيما<br>In what | على ما<br>On what | بما<br>With what | | |
| | which | أيّ<br>(ayu) | In – on - with | في أيّ<br>In which | على أيّ<br>On which | بأيّ<br>with which | | |
| | what | ماذا<br>(matha) | | في ماذا<br>In What | على ماذا<br>On what | مع ماذا<br>With what | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | when | أيّان<br>(Ayyan) | | | | |
| | | متى<br>(meta) | | | | |
| ابتداء الغاية<br>Start of specific aim/ direction/place | which | أيّ<br>(ayu) | من<br>from | من أيّ<br>From which | | |
| | where | أين | | من أين<br>From where | | |
| انتهاء الغاية<br>End of specific aim/ direction/place | which | أيّ<br>(ayu) | إلى – حتى<br>To - until | إلى أيّ<br>To which | حتى أيّ<br>Until which | |
| | when | متى<br>(meta) | | إلى متى<br>Until what | حتى متى<br>Until when | |
| الحالية | Who/ whom | من<br>(men) | في - على - بـ<br>In – on - with | في من<br>In whom | على من<br>on whom | بمن<br>With who |
| | which | أيّ<br>(ayu) | | فأيّ<br>Then which | على أيّ<br>On which | بأيّ<br>With which |
| | what | ما<br>(ma) | | فيما<br>In what | على ما<br>On what | بما<br>With what |
| الخبر<br>Predicate | Who | من<br>(men) | | | | |
| | which | أيّ<br>(ayu) | | | | |
| | what | ما<br>(ma) | | | | |
| العاقل<br>rational | who | من<br>(men) | | | | |
| غير العاقل<br>irrational | what | ما<br>(ma) | | | | |
| الكيفية<br>Description/ Procedure/ Manner | How to | كيف<br>(kayfa) | | | | |
| Quantity/ count | How much/ many | كم<br>(kem) | ف – ب | بكم<br>How much | فكم<br>How many | |

# Appendix E

## E.1 Some results for searching about heaven names in the Quran

| Chapter No. | Verse No. | Matched concept | Similarity percentage | Is match |
|---|---|---|---|---|
| **84** | 12 | أسماء النار السعير | 0.408248290463863 | False |
| **56** | 89 | أسماء الجنة جنة النعيم | 0.866025403784439 | True |
| **25** | 15 | أسماء الجنة جنة الخلد | 0.866025403784439 | True |
| **70** | 38 | أسماء الجنة جنة النعيم | 0.866025403784439 | True |
| **69** | 22 | أسماء الجنة جنة عالية | 0.866025403784439 | True |
| **88** | 10 | أسماء الجنة جنة عالية | 0.866025403784439 | True |
| **53** | 15 | أسماء الجنة جنة المأوي | 0.866025403784439 | True |
| **23** | 11 | أسماء الجنة الفردوس | 0.816496580927726 | True |
| **21** | 101 | أسماء الجنة الحسنى | 0.816496580927726 | True |
| **18** | 88 | أسماء الجنة الحسنى | 0.816496580927726 | True |
| **30** | 15 | أسماء الجنة روضة | 0.816496580927726 | True |
| **13** | 29 | أسماء الجنة طوبى | 0.816496580927726 | True |
| **33** | 47 | أسماء الجنة فضل | 0.816496580927726 | True |
| **2** | 102 | أسماء الجنة الآخرة | 0.816496580927726 | True |
| **43** | 35 | أسماء الجنة الآخرة | 0.816496580927726 | True |
| **56** | 27 | أسماء الجنة يمين | 0.816496580927726 | True |
| **56** | 38 | أسماء الجنة يمين | 0.816496580927726 | True |
| **56** | 90 | أسماء الجنة يمين | 0.816496580927726 | True |
| **56** | 91 | أسماء الجنة يمين | 0.816496580927726 | True |
| **57** | 10 | أسماء الجنة الحسنى | 0.816496580927726 | True |
| **83** | 19 | أسماء الجنة عليون | 0.816496580927726 | True |