Artificial Intelligence Review (2023) 56:6801–6854
https://doi.org/10.1007/s10462-022-10313-2

Arabic natural language processing for Qur'anic research: a systematic review

Muhammad Huzaifa Bashir[1] · Aqil M. Azmi[2] · Haq Nawaz[3,4] · Wajdi Zaghouani[5] · Mona Diab[6] · Ala Al-Fuqaha[7] · Junaid Qadir[8]

Published online: 2 December 2022

Abstract
The Qur'an is a fourteen centuries old divine book in Arabic language that is read and followed by almost two billion Muslims globally as their sacred religious text. With the rise of Islam, the Arabic language gained popularity and became the lingua franca for large swaths of the old world. Devout Muslims read the Qur'an daily seeking guidance and comfort. Though the Qur'an, as a text, is short, there is a huge volume of supporting work filling tens of thousands of volumes, e.g., commentaries, exegesis, etc. Recently, there has been a renewed interest in such religious texts by non-specialists. Many of which were fueled by the recent advances in computational and natural language processing (NLP) techniques. These techniques help the development of tools that benefit common people to gain knowledge easily. This paper surveys the different efforts in the field of Qur'anic NLP, serving as a synthesized compendium of works (tools, data sets, approaches) covering the gamut from automated morphological analysis to correction of Qur'anic recitation via speech recognition. Multiple approaches are discussed for several tasks, where appropriate. Finally, we outline future research directions in this field.

✉ Junaid Qadir
jqadir@qu.edu.qa

[1] Department of Electrical Engineering, Information Technology University (ITU), Lahore, Pakistan

[2] Department of Computer Science, College of Computer & Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia

[3] Jamia Ashrafia, Lahore, Pakistan

[4] Punjab University College of Information Technology (PUCIT), Lahore, Pakistan

[5] College of Humanities and Social Sciences, Hamad Bin Khalifa University (HBKU), Doha, Qatar

[6] Department of Computer Science, George Washington University (GW), Washington, DC, USA

[7] Information and Computing Technology Division, College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar

[8] Department of Computer Science and Engineering, Qatar University (QU), Doha, Qatar

Keywords Arabic natural language processing · Machine learning · Quranic NLP · Religious texts · Classical Arabic

1 Introduction

The Qur'an is the central religious text of Islam,[1] where Muslims believe it was revealed in the Arabic language to Prophet Muhammad (peace be upon him)[2] over a span of 23 years that ended in 632CE, the year the Prophet passed away. The word *Qur'an* appears about 70 times in the Qur'an itself, assuming various meanings. It is a verbal noun of the Arabic verb *qara'a* meaning "to read" or "to recite", an opinion backed by most of the Muslim authorities on the origin of the name Qur'an and also mentioned in Britannica source (bri, 1999).

As a book, the Qur'an is in the Arabic language, and it is typically recited in Arabic. This explains why Arabic is the liturgical language of approximately 1.8 billion Muslims, the majority of which are non-Arabs (about 73%).[3] Knowing the Arabic language helps them understand the true message of the Qur'an. Qur'an holds a lot of knowledge and information in the form of 114 chapters consisting of 6,236 verses in total. There are 157,935 words in Qur'an and out of these 5,277 are unique.[4] These chapters contain teachings regarding daily matters, social dealings, historical events and upcoming events. It is known for being crisp and concise in expression with almost poetic language.

Given the central significance of the Qur'an over more than 1400 years, thousands of scholars dedicated their lives to producing scholarship studying Qur'anic scripture from different angles producing knowledge that easily fills tens of thousands of printed volumes. However, past efforts were manual in nature. With the advances in computational techniques, especially in the field of natural language processing (NLP), NLP is leveraged for facilitating Qur'anic research and studies. Moreover, it opens up avenues for developing new applications that can help those interested in learning and understanding the Qur'an.

1.1 Arabic language and Arabic NLP

Arabic is a Semitic language that first emerged in the first to fourth centuries CE. This is testified by the various Arabic inscriptions found in the region from that era (Al-Azami 2020, pp. 126–129). Modern linguists designate Arabic into one of three main classes: Classical, Modern (or MSA, short for Modern Standard Arabic), and Dialectal Arabic. Standard Arabic is a "prescriptive" term of the language which the early Arab grammarians considered "Classical", while the MSA is a "descriptive" term for realizing Standard Arabic by modern-day Arabs. Vocabulary wise we can safely assert that MSA is a superset of Classical and Standard Arabic (Azmi et al. 2019). It is a superset in the sense that MSA incorporates the names of modern inventions, e.g., radio, computer, while retaining the sentence structure and the vocabulary of its classical and standard form.

[1] This paper is written from the Muslim standpoint, who believe in the Qur'an being a divine scripture revealed to Prophet Muhammad peace be upon him over a period spanning 23 years. This fact may show up occasionally in the written paper, however, it has no bearing on our scientific findings, which should be treated as religiously neutral.

[2] It is a customary practice among the Muslims to invoke "peace be upon him" whenever Prophet Muhammad is mentioned. We opted to drop this salutation to maintain the text's flow as much as possible, with the hope that the Muslim reader will mentally insert these phrases into the text as appropriate. We were guided in this decision by the work of Al-Azami (2020).

[3] https://en.wikipedia.org/wiki/Islam_by_country.

[4] https://www.qurananalysis.com/analysis/basic-statistics.php.

Table 1 The basic Arabic diacritics grouped into three sets

Diacritic Set	Diacritic on letter ب	Name	Pronunciation
Short vowels	بَ	fatha	/b//a/
	بُ	damma	/b//u/
	بِ	kasra	/b//i/
Nunation	بً	tanween fath	/b//an/
	بٌ	tanween damm	/b//un/
	بٍ	tanween kasr	/b//in/
Syllabification marks	بْ	sukon	/b/
	بّ	shaddah	/b//b/

The *nunation* can only be placed at the end of the word. The syllabification mark *shaddah* only occurs with either short vowel or nunation

The Arabic language alphabet consists of 25 consonants and three long vowels. It is written from right-to-left in a complex cursive script that permits a variable degree of stretching or compressing (Azmi and Alsaiari 2014). Most of the characters assume up to four forms per letter, which are contextually shaped. Additionally, the orthographic system uses a total of thirteen different diacritical markings to represent short vowels (a, i, u), and these are placed either above or below the character (see Table 1).

One of the remarkable features of the Arabic language is its expressiveness. For instance, the single surface word فأسقيناكموه, *fAsqynAkmwh*[5] which appears in the Qur'an (15:22),[6] is equivalent in translation to the complete sentence "*and We have given it to you to drink*". Arabic morphology is highly complex and yet systematic. Arabic words are derived from (mainly) triliteral root consonants. Roots hold the base meaning of words. Various morphological operations apply to roots to create patterns that are further morphed into stems (lemmas). Affixes (prefixes e.g., articles, prepositions, and suffixes, e.g. linked pronouns) are appended to lemmas to form lemmatas (Azmi et al. 2019).

NLP is a technique to make computers mechanically process and possibly understand the human (natural) language. With over fifty years of research behind NLP, the field has achieved reasonable maturity. This is particularly true for the English language, for all the obvious reasons. Other languages lag with varying degree. When compared to the English language, Arabic NLP is at least a decade behind. We may attribute this to the rich nature of the Arabic language and its complex grammatical and syntactic structures. Farghaly and Shaalan (2009) detailed some challenges and problems that have to be tackled while performing NLP tasks related to MSA. Most of the work on Arabic NLP involves MSA, the most widespread and literary form of Arabic.

Salloum et al. (2018) noted, a lot of information could be gathered through computational text mining in Arabic literature but the effort in this field is below expectations. This is due to the formidable challenges involved in doing Arabic NLP research. In what follows, we discuss briefly some of the challenges involved in doing Arabic NLP research.

[5] We transliterate Arabic words using the Buckwalter Transliteration system

[6] Chapter 15 verse 22.

إِنَّآ أَنزَلْنَٰهُ قُرْءَٰنًا عَرَبِيًّا لَّعَلَّكُمْ تَعْقِلُونَ

Fig. 1 Qur'anic verse showing the orthography used in the Holy Qur'an

- There is no capitalization in the Arabic script. This makes it hard to identify proper names and further complicates the process of Named Entity Recognition (NER), a basic NLP task.
- The diacritics adds sense and meaning to a word, and the lack of it creates ambiguity. For example, the undiacritized word عقد could be "contract", "necklace", "to make it complex," etc. (Azmi and Aljafari 2018). However, have the word been properly diacritized there would not have been any confusion. Unfortunately, the writing custom in MSA is devoid of any diacritical marking. It assumes the reader can disambiguate the meaning through context. This is a false premise. Azmi and Almajed (2015) presented a sample sentence that requires world knowledge to disambiguate its true meaning.
- Habash (2010) described the problem that is faced during the translation from Arabic as its sentence structure is totally different and the translated sentence has then to be structured properly to make some sense.

1.2 Unique challenges in Qur'anic Arabic NLP

When it comes to Qur'an—in addition to the challenges in doing basic Arabic NLP research highlighted previously—a gamut of new challenges arises. For instance, the text of the Qur'an has its own orthography (spelling convention) which differs even from Classical Arabic. Figure 1 shows a verse of the Qur'an. Anyone who has a slight knowledge of Arabic can easily tell the present-day spelling of the second word should be أنزلناه, the spelling of the rest of the words in the verse are perfectly normal. In fact, all the natural language's orthography evolves over time. For example "cwēn", an Old English word, changed to "quen" in Middle English and later became "queen" (Al-Azami 2020, p. 142). However, when it comes to Qur'an, the Muslim scholars decided to maintain the original spelling used in the second compilation of the Qur'an, which took place during Caliph 'Uthman's reign, a mere 15 years after the Prophet's death (Al-Azami 2020, pp. 95–106). So, even though this orthography differs from the present's, the text of the Qur'an and its orthography has remained unadulterated for the last fourteen centuries.

While the Qur'an forms the bedrock of Islamic law, many legal details are derived not from the Qur'anic scripture, but rather from the utterances and actions attributed to Prophet Muhammad. These prophetic traditions, or hadiths, are narrations originating from the sayings and conduct of Prophet Muhammad. The detail of many of the Qur'anic pieces of knowledge can only be mined in the hadith literature. This is why Muslims believe the hadith complements the Qur'an. For instance, the detail about the practice of praying is only found in the hadith literature. Unlike Qur'an, the hadith corpus is huge and runs into hundreds of volumes. It has an even larger supporting work, e.g., commentaries, biographic material (Azmi et al. 2019). Qur'anic NLP researchers can enrich their Qur'anic NLP research by leveraging the complementary hadith literature and should be cognizant of the fact that ignoring the hadith corpus can lead to erroneous conclusions.

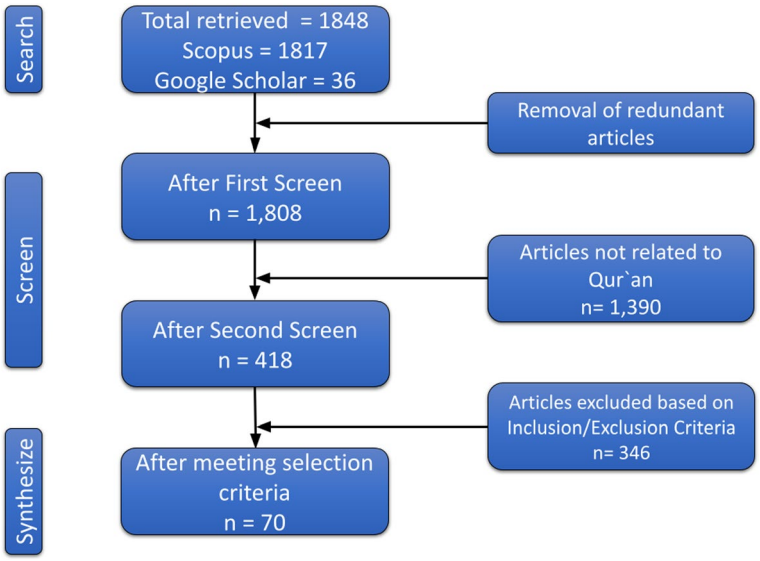


Fig. 2 Flowchart of paper selection

Along with all these mentioned challenges, one of the biggest challenges in working with the Qur'an is that it is considered a divine scripture, which demands extra precaution so that its semantics or information retrieved remains intact (Salloum et al. 2018). Current NLP efforts are focused on overcoming all these challenges using various techniques. Performing NLP tasks in a language with all these problems is not a straightforward task and requires technical as well as linguistic support.

Given the challenges and opportunities of the Qur'anic language, a group of scholars wrote a paper titled "Understanding the Qur'an: a new Grand Challenge for Computer Science and Artificial Intelligence" (Atwell et al. 2010). The authors pointed out that "Understanding the Qur'an" can be considered as a grand AI challenge for various tasks such as reasoning, knowledge representation, and knowledge extraction based on Qur'anic text among other challenges that have to be solved by leveraging the power of the latest NLP techniques.

1.3 Inclusion and exclusion criteria

Papers have been selected based on keywords revolving around Qur'anic NLP. The search criteria uses the NLP keywords combination of "Quran" or "Qur'an" with "NLP" or "Grammar" or "ontology" or "translation" or "search" or "classifier" or "ml" or "knowledge" or "topic" or "corpus" or "question" or "authentication" or "verification" or "crowdsource". The search has been performed on the Scopus and Google Scholar platforms. The papers included meet the following inclusion criteria: (1) articles related to Qur'anic NLP and techniques related to machine learning and deep learning for Qur'an, and (2) articles related to comparisons of different techniques used for a particular Qur'anic domain.

We excluded papers that are not related to Qur'anic NLP, and those not written in the English language. In addition, we ignored incomplete works. Our search returned a total of

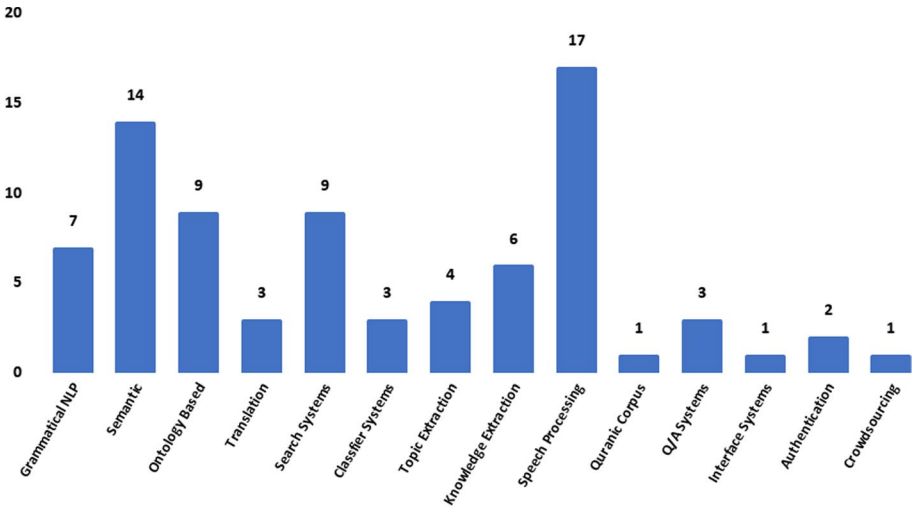


Fig. 3 Distribution of papers categorized according to their topic

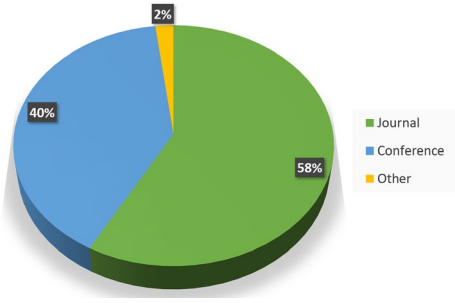


Fig. 4 Distribution of selected publications according to their types

1846 papers. After screening based on inclusion/exclusion criteria (Fig. 2), 70 papers were eventually selected for our survey.

During selection, the focus remained on the works which are closer to the Qur'an instead of Arabic only. NLP experiments usually deal with text-based language processing, however in the survey speech processing based papers for Qur'an have also been considered in the broader spectrum of NLP as it is also the processing of language and may provide useful insights. The papers surveyed can be categorized as shown in Fig. 3.

The surveyed papers were published in different conferences and journals, though the majority were in journals. There is no specific journal or conference which holds dominance in the publication of these papers. The publications have been done during the period of years 2005 to 2022. Figure 4 below shows the percentage share of the category of publication and the quantities of papers surveyed during different years (Fig. 5).

The contribution of reviewed Qur'anic NLP research work is shared from across the globe. The heat map in Fig. 6 shows the spread of research work across the globe for Qur'anic NLP reviewed in this paper.

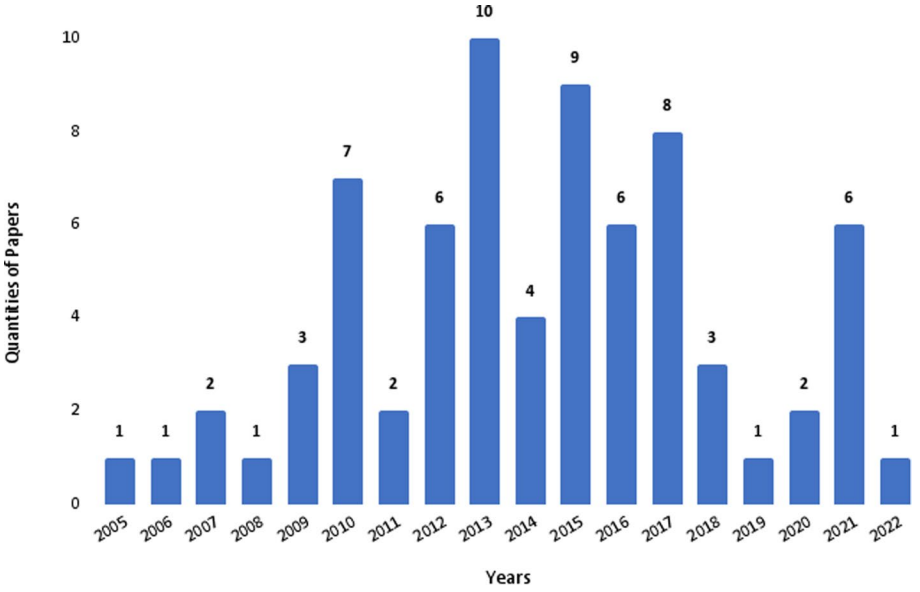


Fig. 5 Distribution of selected publications over years

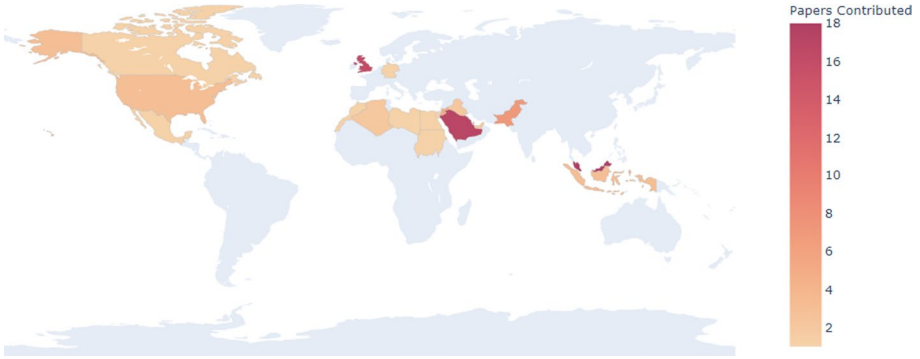


Fig. 6 Heatmap of Qur'anic NLP-related research contributions reviewed in this paper

1.4 Related review articles

The work for Qur'anic NLP is not in limelight currently, but still, a lot of efforts have been made over the past two decades to use the linguistic computation for Qur'an. Kammani and Safeena (2014) have reviewed various tools and techniques in this regard. Beside studying the efforts made in generic Arabic NLP, they have also reviewed the Qur'anic NLP. The review for Qur'anic NLP contributions is ranging from 1997 to 2011. They have discussed the development of Qur'anic corpus, annotations, morphological analysis, semantic searching, use of ontology for Qur'an, grammar tree bank, and translation analysis. However, the review does not provide a variety of work for each category and just discusses one or two

Table 2 Comparison of our work with others

Qur'an NLP work reviewed	Kammani and Safeena (2014)	Atwell et al. (2011)	This paper
Grammatical work	✓	✓	✓
Semantic work	✓	×	✓
Ontology	✓	✓	✓
Translation	×	×	✓
Search system	✓	✓	✓
Classification	×	✓	✓
Topic Extraction	×	×	✓
Knowledge Extraction	×	×	✓
Speech Processing	×	×	✓
Qur'anic Corpus	✓	✓	✓
Q/A Systems	×	×	✓
Interface Systems	×	✓	✓
Qur'an Authenticity	×	×	✓
Crowdsourcing	×	×	✓
Review Timeline	1997–2011	2004–2011	2005–2021
Focus	Generic & Qur'an	Qur'an	Qur'an

tools or techniques. There are other domains such as the use of NLP and speech recognition for Qur'an recitation, question answering tools for Qur'an, knowledge extraction and topic modeling, which have not been highlighted.

Zaghouani (2014) presented a survey of freely available corpora for the Arabic language with a limited number of papers related to the Qur'an and the Hadith. Atwell et al. (2011) presented their own contribution towards Qur'anic NLP. The paper is focused on how their work can be useful in the correct interpretation of Islam in the Western world. Qur'an-related NLP contributions described by them include open-source Qur'an corpus, conceptual search tool for Qur'an, grammatical analysis, classification of verses and chapter, ontological work and online tools. The paper does not cover much detail of each contribution.

Though not directly related to Qur'anic research, Azmi et al. (2019) provides a comprehensive survey of NLP-based works on the complementary hadith literature. We note that the hadith literature, along with Qur'an, is one of the two canonical sources of Islamic knowledge.

Our work is focused on Qur'anic NLP instead of generalized Arabic NLP, as this would help in understanding the current status of Qur'anic NLP and will help researchers to contribute further with the help of already built applications. Techniques adopted for doing the Qur'anic NLP tasks have been discussed. The paper also covers the tools that have been used in various Qur'anic NLP tasks. The comparison of this paper with previous such reviews is shown in Table 2

1.5 Contributions of this paper

This paper aims to review the Qur'anic NLP research work performed including the techniques and tools used in them. Surveyed work is related to the applications and systems

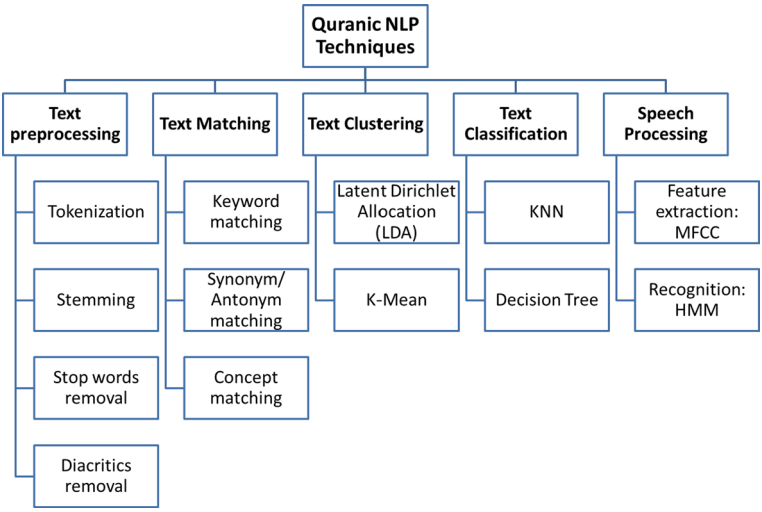


Fig. 7 Techniques used for Qur'an NLP tasks

developed for Qur'an based on NLP. The paper will highlight various challenges that have been overcome related to Qur'anic NLP, major contributions, and the areas that still require attention. The main purpose of writing this paper is to consolidate the current status of Qur'anic NLP works and help the researchers to work ahead of that. The papers have been searched based on the following survey objective:

1. To study and survey the current research work related to Qur'anic NLP.
2. To study the techniques, tools, and resources being used for Qur'anic NLP.
3. To find the limitations and pitfalls of Qur'anic NLP works.

1.6 Organization of this paper

The rest of the paper is organized in the following way. Qur'anic NLP techniques are described in Sect. 2. A broad survey of Qur'anic NLP-related works is provided in Sect. 3. A discussion on various tools and resources available for Qur'anic NLP research follows in Sect. 4. The various caveats of doing AI-based NLP research for Qur'anic research and potential pitfalls are discussed in Sect. 5. Open issues and future research directions are identified in Sect. 6. Finally, the paper is concluded in Sect. 7.

2 Qur'anic NLP techniques

In NLP, different techniques are involved to gain some useful output from the input data. Some techniques are basic and are used in almost all applications while others are task-specific. Figure 7 shows the overview of techniques involved in the Qur'anic NLP tasks and they are further explained in each subsection. In the remainder of this section, these are discussed in more detail.

2.1 Text preprocessing

Text preprocessing is one of the basic steps that is used in almost all textual NLP tasks. Preprocessing involved in the NLP tasks of the Qur'an mostly begins with tokenization. The given corpus from the Qur'an may consist of multiple chapters, divisions, or verses. One verse is like a sentence. These verses are tokenized into words. In Arabic, a particular word may consist of multiple units, so depending on the requirement, these words can be further segmented into morphological segments. In certain applications, such as question-answering platforms or search systems, Qur'anic words are stemmed to their roots. Like English, Arabic also contains many stop words. Stop word removal is essential in tasks like topic modeling of the Qur'an and finding similar verses. For the help of Non-Arab Muslims, diacritic marks are present on the words of Qur'an. These marks help as a guide for the reciters to avoid mistakes and confusion. However, these marks can increase the complexity while processing the text through the machine. In many tasks, it has been observed, the diacritics are removed from the words in the preprocessing stage. One of the reasons is that if the experiment requires finding a particular word, it might fail as the same word may have different diacritics depending on the usage in the sentence. Although, it is not essential to remove diacritics always and they could be used as features for some experiments too where required.

2.2 Text matching

Text matching technique has been widely used in search systems developed for Qur'an. Text matching can be used in multiple forms. In some applications, the exact words of the user query are matched with that of verses. In other applications, the query words and the words of verses may be converted to their roots so that they could be matched to any form they have been used in. This technique increases the search window. Advanced forms of text matching, where the matches are not only made based on words, but rather concepts behind those words are matched, can be used in semantic tasks related to the Qur'an. For example, a simple search may only focus on some search words and retrieve the results and they might fail to provide correct intended results. Here, the search window size can be further enhanced using synonyms or antonyms of a particular word. Retrieving concepts from any text cannot be done by simple word matching. Conceptual results can be retrieved using ontology where the matching of words is performed at different hierarchical levels. Although text matching may seem to be a very basic technique, if used properly, it can help in achieving great results.

2.3 Clustering

Clustering is an unsupervised machine learning technique that is used to cluster similar kinds of data. In unsupervised learning, labels are not available. For the Qur'an, similar kinds of verses are grouped for the categorization and knowledge extraction tasks. There are various algorithms for clustering but for topic modeling in particular Latent Dirichlet Allocation (LDA) is the most widely used. LDA calculates the probability of words distributed over a topic and then allots multiple topics to each document (Alhawarat 2015). Topics discussed in Qur'an can also be listed, by using the clustering technique on the Qur'anic verses. Besides, LDA, K-means algorithm is also applied for the categorization of the Qur'an verses. K-means is a clustering algorithm in which given verses are clustered into a known

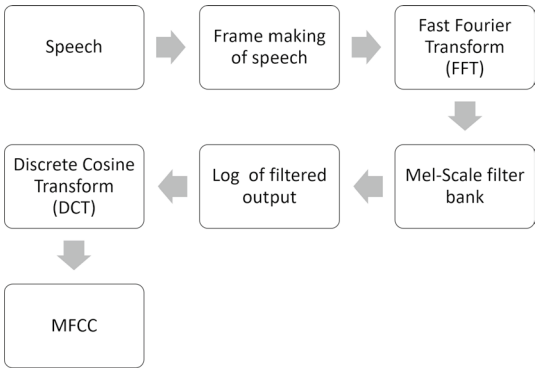


Fig. 8 Block diagram of MFCC feature extraction from speech

number of groups, depending on the distance of words in verses from the centroids of each group. K-means algorithm is simpler in implementation as compared to LDA.

2.4 Classification

Contrary to clustering, classification is a supervised machine learning technique that is also used for the topic modeling of the Qur'an and knowledge extraction. In classification, data is already labeled. K-Nearest Neighbours (KNN) is widely used in classification tasks. In KNN, there is a set of verses which already labeled with particular topics. The verses under test are then allocated a topic by finding their similarity with labeled verses using a distance formula. Decision tree is the other classification model. The model is used in the application of knowledge extraction and question answering for Qur'an (Mohamed et al. 2015). In the decision tree algorithm, training data is classified based on the input features. The testing data is then allotted labels based on the rules developed by the decision tree during training.

2.5 Speech processing

Speech processing has been used in multiple tasks related to the correct recitation of the Qur'an, as discussed in Qur'anic NLP works section. Speech processing helps in conveying the natural sense of spoken languages. For Qur'an, it is essential to recite with certain rules. In speech recognition, Mel-Frequency Cepstral Coefficient (MFCC) is used for feature extraction of the speech. MFCC is designed to replicate the human auditory system artificially (Chakroborty et al. 2007). The feature extraction using MFCC is shown in Fig. 8.

Hidden Markov Model (HMM) is used for training and testing of speech data. HMM-based recognizer takes feature vectors created by MFCC as input and then uses Bayesian probability to recognize words from phonemes trained over multiple pronunciations of each word (Gales and Young 2007). HMM is a state machine where each state in a particular time is associated with a specific phoneme and a particular word is obtained by observing this sequence of phonemes (Dimitrakakis and Bengio 2011).

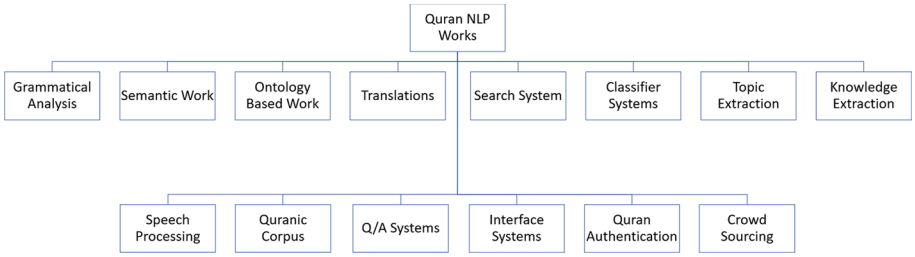


Fig. 9 Categorization of NLP works

Table 3 Comparison of treebanks

Treebank	Dependency	Feature	Traditional
Penn	No	Yes	No
Prague	Yes	Yes	No
Columbia	Yes	No	Yes
Qur'an	Yes (hybrid)	Yes	Yes

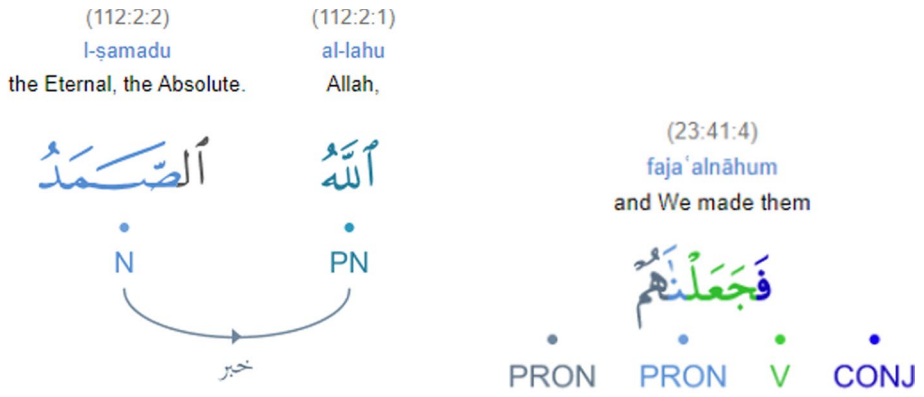
3 Qur'anic NLP works

Quranic NLP tasks are reviewed in this section. Figure 9 shows the categorization of tasks.

3.1 Grammatical NLP analysis

Arabic grammar can be understood in terms of morphology and syntactical analysis. The syntactic analysis, along with morphology in NLP has to cater to all the rules, such as the use of the correct form of the word, placement of a word in the correct place within a sentence, and relating nouns verbs, and particles properly. For this purpose, various Arabic tree banks have been developed. To study the Qur'anic grammar, Dukes et al. (2010) have developed Qur'anic Arabic Dependency Treebank (QADT). Treebank is a collection of manually annotated text. QADT is part of Qur'anic Arabic Corpus (QAC) (Dukes 2009–2017). Morphological analysis for all the 77,430 words of the Qur'an has been completed with the help of a collaborative approach. However, syntactic relation analysis is yet in progress and has covered 11,000 words so far. Table 3 shows the comparison of other general Arabic treebanks with Qur'an dedicated QADT:

Morphological annotation involves the segmentation of each word into morphemes and assigning part of speech to each of them. The morphological annotation adopted in QADT for the Qur'an is described by Dukes and Habash (2010). A three-staged systematic method is adopted for morphological tagging which includes automatic tagging, two-pass manual verification, and online collaborative annotation. Automatic tagging is done by using the existing Buckwalter Arabic Morphological Analyzer (BAMA) aided by a custom tool. In manual tagging, one of the annotators verifies the automatic tags and the second one reviews the text after her/him. Finally, the accuracy is further improved by an online



(a) Dependency graph for verse showing syntactical structure (112:2)

(b) Morphological segmentation for word (23:41:4).

Fig. 10 Visualization of Dependency graph and Morphological analysis of Qur'an (Dukes 2009–2017)

collaborative platform. The automatic algorithm provides analysis for 87% words of the Qur'an; the remaining 13% were not catered because of the low vocabulary of BAMA. For 87% words, automatic annotation provided 77% accuracy and 83% recall.

Traditional Arabic has complex functional inflection. For example, a noun can be written in different ways depending on its position against its verb. Also, a single word may consist of multiple morphological units syntactically related to each other. Dukes et al. (2010) have explained the process of syntactic annotation and key guidelines for annotations of Qur'anic verses based on QADT. Guidelines are important because Qur'an is a divine scripture. An iterative approach is adopted to annotate the corpus similar to that of morphological annotation. Initially, a rule-based dependency parser having an F-measure of 78% is used. The next step is to perform a manual check of annotation for corrections using existing traditional grammar rules and annotation guidelines. Interested volunteers regularly participate in the online annotation. Annotation guidelines are added and updated based on this online collaborative work. There are multiple guidelines such as sentence structure based on subject, verb, and object, use of empty and hidden nodes for reconstruction, and prepositional phrase attachments. The supervisors have a role to veto any decision as they are more knowledgeable and they see if the annotation done is aligned with Qur'an or not (Dukes et al. 2013).

A visualization tool can be a great replacement for traditional methods and help Arabic linguists and Qur'anic students understand traditional Arabic grammar more easily. Dukes et al. (2010) report the capability of graphical dependency graphs in QADT to present the syntactical relationship of words with a sentence. Each word may consist of multiple segments. To perform this computational analysis, graphs have been developed which connect different words within a sentence or with words in other sentences. This technique allows understanding the relation of different segments, words, and phrases in the Qur'an. Detailed morphological visual analysis of verses is also provided along with a color-coded grammatical analysis of each morpheme. POS tagging and morphological annotations provide high accuracy under random sample testing. Figure 10 shows the visual representations for morphology and syntactic analysis.

Fig. 11 The full i'raab of Surah 73 verse 20 from the Holy Qur'an. Source: Mannaa et al. (2022)

Bentrcia et al. (2018) have studied the conjunctive phrases in Qur'an based on the 'AND' conjunction and have shared interesting findings. Three different cases are recorded with 'AND' conjunctive phrases which are; words occurring in specific order only once, words occurring in specific order multiple times, and words occurring in a different order one time or multiple times. The corpus used for the project is the QAC and a combination of the statistical and grammatical methods have been adopted to mine the required conjunctive phrases. There are many cases in which words are combined by 'AND' but in the study, ten patterns that consist of only nouns, pronouns, and adjectives combined by 'AND', are considered. The results and analysis reveal that words occurring in some specific order are based on some logic. For example, 'Isaak' is always mentioned before 'Jacob' because the former was the father of the latter and 'East' is always mentioned before 'West' because in the time frame east comes before the west. The order of the words in conjunctive phrase hold significance based on the context of the verse. Pointwise mutual information is also recorded for the words joint by the 'AND' word which shows that words occurring multiple times together have a higher association.

ElAffendi et al. (2021) recently used a neural network based approach to predict the morphological patterns and POS tags for Arabic. Galois Power-of-Two (GPOW2) calculates the real-time embeddings for character, word, and sentence in parallel. Each word is represented as a polynomial of power of two. The QAC is used as a dataset that holds rich morphological and syntactic annotations. The model works by computing the contextual embeddings of a word at input stage. GPOW2 representations of these embeddings are fed to the neural network which then tries to predict the tag for the target word. The test accuracy for POS tags prediction is 98.8%, which is impressive when we consider the complexity of Arabic morphology.

In linguistics, the process of identifying each word's function in an Arabic sentence is known as علم الإعراب or just إعراب (i'raab) in Arabic. This is accomplished by adding the appropriate diacritical mark and an appropriate justification or reasoning-previously known as the end-case analysis-at the end of each word. As a result, syntax parsing is one of the most crucial aspects of Arabic because a weak or incorrect i'raab causes a misunderstanding of a sentence's precise meaning. In light of the fact that the Holy Qur'an is written in

Table 4 Qur'an grammatical NLP works table

Sr.	Research	Issue addressed	Technique applied/contribution
1.	Dukes and Buckwalter (2010)	Lack of manually verified Qur'anic corpus and treebank	Development of QADT and Qur'anic Corpus
2.	Dukes and Habash (2010)	Lack of morphological annotation of the Qur'an	Systematic, collaborative approach for morphological annotation of the Qur'an
3.	Dukes et al. (2010)	Lack of syntactic annotation of the Qur'an	Systematic, guideline based and collaborative approach for syntactic annotation of Qur'an
4.	Dukes et al. (2013)	lack of supervision of syntactic work for the Qur'an	Inclusion of Supervisors to monitor the annotation
5.	Dukes et al. (2010)	Lack of user friendly interface to understand Qur'an grammar	Dependency graph for verse structure and Color coding for morphology
6.	Bentrcia et al. (2018)	Understanding logic in placement of words around conjunctions in Qur'an	Study of placement of nouns, pronouns and adjectives occurring with 'AND'
7.	ElAffendi et al. (2021)	Prediction of POS for Qur'an	Use of GPOW2 for context-based POS prediction using neural network
8.	Mannaa et al. (2022)	I'raab of Arabic sentences	Enhanced context-free grammar

Arabic, it is crucial that we understand its meaning in its entirety. Mannaa et al. (2022) devised an enhanced context-free grammar (eCFG) that covers all the rules taught in Saudi school's grammar textbooks. The eCFG eliminates the need for specialized grammar, e.g., link grammar, to resolve complex cases involving dependencies. The authors tested their system on 300 sentences and reported an accuracy of 88.33%. Figure 11 shows a sample i'raab of a verse from the Holy Qur'an.

Table 4 summarizes this section.

3.2 Semantic technologies

Sherif and Ngonga (2015) have developed a database that contains semantic datasets in 43 languages for the Qur'an. Tanzil is a web-based resource of multi-language Qur'anic translations for the dataset and QAC provided the morphological data for each word. The data has been presented in the Resource Description Framework (RDF). Out of 43 languages, only three translations along with the original Arabic dataset comply with the RDF and NLP Interchange Format (NIF). NIF achieves inter-operatability between NLP tools, language resources, and annotations. RDF consists of four main classes of Chapter, verse, word, and lexical term which give a hierarchical structure to the whole dataset. The dataset has been linked with multiple online platforms for use. It helps in data retrieval from the Qur'an in multi-languages. Using the SPARQL, the user can also perform morphological queries on data. Other useful applications include information aggregation and finding multiple occurrences of a particular text in the database.

Al-Khalifa et al. (2009) have proposed a web-based framework under the title "SemQ" which performs semantic opposition analysis on Qur'an using NLP and semantic web technologies. In semantic opposition analysis, opposites of a word or phrase are searched based on their characteristics. SemQ takes input in the form of verse and returns it provides an output which is a list of words that are semantically opposite to each other. Ontology based on Greek New Testament and SemQ Tool work together under the framework. After NLP preprocessing, the algorithm starts working by checking if the two words under consideration belong to the same category. If only one property of words differs, they are declared absolute semantic opposite; if more than one properties differ then the scale of opposition is calculated. The SemQ team aims to build prototypes in the future for the complete Qur'an and compare the results with traditional approaches.

Afzal and Mukhtar (2019) have tried to address the problem associated with the keyword search system for Qur'an. Many times, intended results are not achieved as keyword search does not cater to conceptual or semantic analysis for the query. They have presented a semantic and lexical-based search system to overcome the issue. Database with the name of Qur'anic English WordNet (QEWN) is developed. In this database, all the words in the English translation of the Qur'an are populated along with their semantic information. Vocabulary of Qur'anic Search (VQC) stores various concepts included in Qur'an. Concepts of sense and Synset is utilized in QEWN. Sense means a particular meaning of a given word and Synset is the collection of synonyms of that word in a particular sense. Upon search of a particular word, all the semantically related verses are returned. Even if the search word does not exist in translation, the tool searches for the related concepts using synonyms. The tool also performs a unique search of Islamic concepts such as the "pillars of Islam" which will return all the verses related to the five pillars of Islam. This semantically rich tool has been able to achieve 88% recall for the words which are directly

present in translation and 59% for the words that do not appear directly but are searched using different synonyms in a particular sense.

Shoaib et al. (2009) have also tried to resolve the problem which is associated with keyword-based search in which either exactly required verses are not retrieved or irrelevant results are returned They have also utilized WordNet database to perform a semantic search on Qur'an. Topic search is a two-step process. In the first step, desired meaning of the query word is interpreted while other senses are ruled out. Afterward, the topic is searched using that word or synonyms of that word. These synonyms are categorized as an exact synonym, close or strong synonyms, and weak synonym. The systematic algorithm returns verses in the priority list. The model has been implemented on English translation of *Surah Al-Baqarah* (Chapter 2) using SQL and VB.Net. The model is also able to achieve great results by retrieving 80% more relevant results as compared to other models. It is also able to minimize the retrieval of irrelevant results.

Malhas and Elsayed (2020) built a Qur'anic test collection, AyaTEC. The objective is to provide a benchmark for evaluating potential QA systems, rather than building a QA system. AyaTEC is a verse-based question answering on the Holy Qur'an, which includes 207 questions with their corresponding 1726 answers. The questions cover eleven topics of the Holy Qur'an that matches the need of curious and skeptical users. The latter is a user seeking answers from the Qur'an to questions that may include controversial or undermining questions. The answers to the questions (each represented as a sequence of verses) were exhaustive—i.e., all Qur'anic verses that directly answered the questions were exhaustively extracted and annotated. AyaTEC covered factoid and non-factoid questions.

Alqahtani and Atwell (2015) have reviewed various techniques to perform a semantic search on Qur'an and have proposed their own framework. Search techniques fall into two types. One is semantic search in which the concept-based search is applied and the other one is keyword-based search in which direct search is performed using some particular word without catering to the concept. The Qur'anic semantic search methods can be categorized as ontological search (concept-based), synonym-set search (based on synonyms of query word) and cross language information retrieval (CLIR). In the keyword search, if any of the query words are matched with any word of the verse, the verse is retrieved. Preprocessing on words is performed with the help of morphological analysis. Keyword-based search also utilizes the chatbot technique in which some important words are chosen from the whole query and a search is performed for those words only. However, keyword-based search frameworks are deficient in providing relevant results and many times provide irrelevant verses. On the other hand, semantic search does not cover all the concepts of the Qur'an. Keeping these limitations of both techniques in view, a new semantic framework is proposed called Qur'anic Semantic Search Tool (QSST). When a query is entered, it is preprocessed and results are retrieved separately for both semantic and keyword search. Redundant verses are eliminated and finally, verses with the highest ranking as scored by QSST are returned to the user.

Yunus et al. (2010a) have performed a semantic search using CLIR and showed the results visually using a space tree model. The system supports search in Malay, English or Arabic. Each word in the query is stemmed to its basic form along with its synonyms. This increases the search range and can help in the retrieval of more results. Finally, the results are retrieved and verses IDs are displayed in the form of a tree structure where the system is further integrated into the speech query (Yunus et al. 2010b).

Some of the modern semantic searching techniques for Qur'an are discussed here. Ensaf and Eyad (2022) have utilized embedding matrix for semantic-based search from Qur'an. The method creates an embedding matrix that is trained on Qur'anic and classical Arabic

corpora. This creates feature vectors for the verses based on words. When a query is made, a feature vector for the query is also made, and then cosine similarity is used to find the semantically closest results from Qur'an. They achieved a recall of 72%. They have suggested experimenting with Doc2Vec instead of word-based vectors as they might improve results. Menwa et al. (2021) used the technique of Doc2Vec to find similar verses in the Qur'an. Each verse is vectorized and a similar verse is found by finding the vector in a similar direction. They used cosine similarity to find similar verses. They achieved an accuracy of 76%. The method of Doc2Vec can be used in finding semantic-based searches because this would help in finding the results based on better semantics as each verse would be a vector and the cosine similarity would be providing a degree of similarity between the query question and verse.

Muhamad Fahmi et al. (2020) have used the text summarizing technique based on the text rank method which focuses on the most important words in a document and returns the summary. Initially, the semantic search is done for the query, and results are retrieved using word2Vec and cosine similarity. The retrieved documents are then summarized and given as output. These documents are not the actual Qur'an but an encyclopedia of the Qur'an in the Indonesian language. The idea behind this technique is to provide a summary of relevant documents to the user without changing the context. Ali and Maged (2020) also used the word2vec technique and proposed that their methodology can be used in multiple applications. They addressed the need to cater search system for Islam-specific knowledge. For this purpose, an embedding of words based on the Qur'an and Hadith was designed. The skip-gram-based model was made as they found it better for domain-specific vectorization. They also tested the model with different Islamic terminologies which were closer in context and the results were satisfactory. The technique adopted seems to be fruitful for the semantic query system as embedding made solely on Islamic resources will be able to answer queries in the right context.

Faiza et al. (2021) have used query expansion technique for semantic Qur'anic searching. The ontology of words and meanings is used where when a query is made, each term of the query is further enriched and a set of new queries are formulated. These queries then search for all the possible semantically closest results for the user. A precision of 70% was achieved with this algorithm. Muhammad et al. (2021) proposed a framework for semantic graphs for the Qur'an. The framework would be able to answer the queries considering the semantics of the query. For this purpose, they proposed a system where word dependency would be used within a verse that would show the relationship between words. Along with it, POS tagging would also be used. A semantic relation consisting of subject-predicate and object would show the purpose or logic behind each verse. Using these word dependencies, POS tags of words and semantic graph rules would be generated that would help in finding better results for the queries. A summary of this section is provided in Table 5.

3.3 Ontology based technologies

Ontology is a hierarchical dictionary of concepts. Since machines cannot understand the concepts like humans, so we need to provide them with an easy way to understand the concepts. Ontology is also used for Qur'an to explain the various concepts easily. Al-Yahya et al. (2010) have made an effort to contribute to the Arabic computational lexicon using the approach of ontology. A lexicon can be defined as the vocabulary of a language where each word is set accompanied by a few other words which define its properties. 59 nouns related to the semantic field of "Time" are chosen in the project

Table 5 Semantic work for Qur'an

Sr.	Research	Issue addressed	Technique applied or contribution
1.	Sherif and Ngonga (2015)	Semantics of Qur'anic Words	Development of the Qur'an semantic database and integration with online platforms
2.	Al-Khalifa et al. (2009)	Semantic opposition analysis of the Qur'an	Development of "SemQ" to perform semantic opposition analysis of Qur'anic verses
3.	Afzal and Mukhtar (2019)	Lack of semantics in keyword-based search for Qur'an	Development of QEWN database and VQC (Qur'anic vocabulary) search system for Qur'an based on synonyms of words
4.	Shoaib et al. (2009)	Lack of semantics in keyword-based search for Qur'an	Retrieval of English verses of Al-Baqarah chapter, based on various degrees of synonyms.
5.	Alqahtani and Atwell (2015)	Study of shortcomings of existing Qur'an search systems	Development of QSST to perform semantic + keyword-based search
6.	Yunus et al. (2010a)	Semantic search of the Qur'an	Development of CLIR system for Qur'an
7.	Yunus et al. (2010b)	Lack of user-friendly semantic search system	User-friendly tree structure based output along with speech query
8.	Ensaf and Eyad (2022)	Enriched semantic results	Embedding matrix based search system
9.	Menwa et al. (2021)	Enriched semantic results	Doc2Vec based vectorization of verses
10.	Muhamad Fahmi et al. (2020)	Lack of summarized answers related to Qur'anic topics	use of text rank algorithm along with semantic cosine search
11.	Malhas and Elsayed (2020)	Lack of benchmark dataset for Qur'anic QA system	AyaTEC dataset and evaluation of a QA system that returns verses
12.	Faiza et al. (2021)	Semantic search of Qur'an	Use of query expansion technique with ontology
13.	Muhammad et al. (2021)	Semantic search for Qur'an	Development of semantic graph framework
14.	Ali and Maged (2020)	Semantic system for Islamic knowledge	word2vec embedding on Islamic specific vocabulary

from the Qur'an. Component analysis of the words belonging to the same semantic field and having similar contexts helps in differentiating them from one another. The team adopted Unified Process for Ontology (UPON) which is a unified software development process-driven methodology for the development of the ontology. Words were classified in a hierarchy having general concepts leading down to more specific ones. There were 18 conceptual classes defined for the time of which eleven were specific to the time field. It can provide synonyms as well as antonyms based on the component analysis. The model was tested for 31 other time words and also words from the human semantic field, given in Qur'an.

Faiza et al. (2021) developed an ontology for Qur'an to find the semantic results for queries. The ontology was developed based on the words of the Qur'an along with their meanings. These meanings are also linked to the concepts. The base of the design is that terms of the query would be searched in ontology and the relevant concept would be retrieved to maximize the search results.

Iqbal et al. (2013) have pointed out the limitations of already existing Qur'anic ontologies and also developed a new ontology for the 30th *Juz* (Division) of the Qur'an to address those limitations. The list of limitations, highlighted by the team while studying previously developed ontologies is described next.

1. Ontology developed for just one topic, for example, Salah.
2. Limitation of ontology to answer all the questions related to a topic.
3. Limited concepts inclusion in the ontology.
4. Linking different concepts related to verses but not catering to their contextual meanings.
5. Methodologies adopted to develop the ontologies do not follow some systematic approach.

The ontology developed by Iqbal et al. (2013) is sourced from an authentic Qur'anic corpus. Ontology is capable of providing the contextual meaning of verses based on Hadith and Tafsir. Ontology has been developed by merging different ontology methods. It provides details of *surah* (chapter) included in the *Juz* (division). Ontology is developed for English and Malay translation. Two ontology development methods have been adopted in the project, namely, Gruninger and Fox's methodology and METHONTOLOGY methodology (Fernández-López et al. 1997). The Ontology developed was able to provide the correct answer to the queries. The framework can be adopted in many semantic and online Qur'anic applications owing to its flexible design.

Khan et al. (2013) have proposed an ontological approach to perform a semantic search for the Qur'an. Ontology for the animal domain has been developed to avoid any conflict regarding divine concepts. Developing animal ontology for 167 animals, described in Qur'an, is also not an easy task as the context has to be specific as per Qur'an. The team has selected the English translation. To improve the results further, the team developed a separate ontology on scientific facts or other related information about animals and then linked it to the original Ontology of the Qur'an. In this way, the actual concept of the Qur'an is not mixed with other data. To perform the Query, SPARQL query language is used. The framework provides great results. It has been proposed to develop Qur'anic WordNet that will further help to perform a better semantic search on Qur'an as the currently available WordNets are generic and do not suffice the requirement of specificity of Qur'anic concepts. To enhance the framework further for the whole Qur'an and other Islamic literature, the team has proposed a simple workflow.

The query will be passed through Qur'anic WordNet and ontological concepts will be linked to the words. Finally, the answer will be retrieved.

Ali and Ahmad (2013) have also used ontological methods to represent the Qur'an under a thematic structure. In the Qur'an, the same concept is described in more than one verse and *surah*. The team has worked to bring all the verses belonging to one theme under one umbrella. To perform the thematic classification, the team has selected Syammil Al-Qur'an Miracle, and only two themes are selected which are *Akhlaq* (Manners) and *Iman* (Faith). Classes are defined where each class has a subclass. For example, Iman is the main theme, and belief in Allah is its sub-theme. Each sub-theme can have different divisions, chapters, and verses. Themes are developed from the reference book and also by interviewing expert scholars. The ontology is developed in the Malay language. and validated by seven experts of the Qur'an. The ontology-based thematic applications can be very helpful in understanding concepts of the Qur'an in a better way.

Ta'a et al. (2017) also developed theme based ontology named Al-Quran to perform the semantic queries. Ontology is developed using Protégé-OWL. The ontology does not cover the whole Qur'an, but only three main classes which are; Allah, Angels, and Unseen. For these classes, sub-themes are developed with the help of Islamic scholars to keep the original meaning intact. The Qur'anic data is saved in a database on which SQL queries can be performed to retrieve results. The graphical user interface (GUI) of ontology allows the user to select the main theme and two subthemes. This helps the user to search for the intended query more easily.

Alqahtani and Atwell (2016) have reviewed the already built Qur'an ontologies along with Qur'anic Search applications and tools. Applying semantic search to Qur'an is not an easy task. One verse may contain multiple themes or a single theme may exist in multiple verses. Different search tools that have been developed try to address different challenges. Still, there are a few limitations in the proposed semantic search tools, which are described next.

1. Unavailability of a solution if the terms of the query do not match the concept of ontology.
2. Limited vocabulary of Arabic WordNet.
3. Incomplete link development of verses with their concepts.

Similarly, the authors have performed research on the available Qur'anic ontologies. Most of these ontologies revolve around verses' similarity and relationship, animals, or some specific Qur'an topic. The team has proposed a new search tool called Arabic Qur'anic Semantic Search Tool (AQSST) by combining four different ontologies. The database consists of original Qur'anic text along with eight different English translations. It also contains Tafsir which is an explanation of concepts of verses. NLA performs parsing and semantic tagging to the query words and then a concepts-based search is performed using SRM. If the search is unsuccessful, then KSM is applied. Finally, suitable results are retrieved based on the score provided by SR. The tool has tried to overcome the limitations and challenges which are highlighted at the start.

Ahmed and Atwell (2016) have investigated the practical applicability of performing a semantic search for abstract concepts of the Qur'an based on different ontologies merged. There are different methods to merge the ontologies such as similarity measure, heuristic method, and methods involving semantics and syntax. The three ontology methods applied in this ongoing project are:

Table 6 Different ontologies and their features

Ontology	Feature
Qur'anic Topics	1100 concepts linked to all verses. It is based on the scholarly book Mushaf Al Tajweed
Arabic Qur'an Corpus	300 concepts and 350 relations from the Qur'an. It is based on "Tafsir Ibn Kathir"
Qur'an	1050 concepts and 2700 + relations to Qur'anic verses

1. Extract abstract concepts with the help of a domain expert.
2. Semi-automatic extraction of concepts from the original text source.
3. Reuse the existing partial ontologies.

Initially, three different ontologies are merged. These ontologies are combined based on the Jaccard similarity which measures the similarity between sets of data. The other method adopted is using a PROMPT tool which takes two ontologies as input and then returns one merged ontology. Both methods provide satisfactory results. The first method gives a correct resolution of results up to 82% while the second provides 85%. The approach in the future is to complete all three methods mentioned at the beginning and create a single ontology that will be sufficient to perform the semantic search on Qur'an.

Alqahtani and Atwell (2016) have discussed the advantages of aligning and merging different ontologies. Only a few among the various ontologies for Qur'an discuss all the concepts. The three most common ontologies with their features are discussed in Table 6.

To merge these ontologies first of all they need to be normalized so that all of them have the same format. After that, the same concepts are aligned based on string matching and semantic matching. There is also a structural approach in which entities are matched based on an ontology graph. The advantage of merging different ontologies into one is that more knowledge is gathered in one place and it can enhance the search results of topics for Qur'an.

The section is summarized in Table 7.

3.4 Translations and Qur'anic NLP

Kammani and Safeena (2013) has discussed the challenges related to Qur'anic translation and then proposed a conceptual method to overcome them. Very little effort has been made to translate the Qur'an based on knowledge. Many researchers have pointed out various challenges regarding translation. Morphological and lexical challenges are the most common among all challenges. Many native Arabic speakers feel challenged when identifying the exact meanings of some words or sentences from the Qur'an. However, a knowledge-based translation developed with the help of new technologies can provide exact meanings for such words or sentences. In the current era, a lot of effort is being put in Qur'anic Arabic and modern Arabic computation research work. Three research gaps were found in these works. Firstly, ignorance to the context and chronological order of the Qur'an; secondly, the inability to answer complex questions; and finally, existing work does not offer complete capabilities of search and analysis. To address these gaps a translation system is proposed in which Qur'an and authentic Hadith will be the knowledge base and at the top

Table 7 Works based on Qur'anic ontologies

Sr.	Research	Ontology domain	Technique applied/contribution
1.	Al-Yahya et al. (2010)	Time domain	Hierarchical system to find synonyms and antonyms of words in Qur'an
2.	Iqbal et al. (2013)	30th Division of the Qur'an	Ontology provides authentic contextual meaning of verses. Translation available in English and Malay
3.	Khan et al. (2013)	Animals	Development of framework to perform semantic queries related to animals in Qur'an
4.	Ali and Ahmad (2013)	Manners and Faith	Development of thematic hierarchy based on an authentic source
5.	Ta'a et al. (2017)	Allah, Angels and Unseen	GUI based thematic search system for Qur'an
6.	Alqahtani and Atwell (2016)	Complete Qur'an	Merger of existing ontologies and development of AQSST search tool
7.	Ahmed and Atwell (2016)	Various concepts but not all	Development of ontology by the merger of existing, extracting original text from the Qur'an, and help from domain experts
8.	Alqahtani and Atwell (2016)	–	Proposed the idea of existing complete Qur'an ontologies to widen the availability of knowledge
9.	Faiza et al. (2021)	Ontology to find semantic results	Development of ontology based on Qur'anic words and meanings

of it will be an expert engine that will be responsible for the content, context, and chronology of the verses.

Putra et al. (2017) have reviewed concepts related to text mining, searching and question answer (SQA) application, and Indonesian Translation (ITQ). The review is conducted in three phases which are planning, conducting, and reporting. Many people search for solutions to their problems from the Qur'an in the Indonesian language. SQA is a search engine developed based on ITQ to provide information to people in the Indonesian language. Discovering Source Answer (DSA) is the component of SQA whose function is to perform sentence detection, POS tagging and text processing, etc. The other component interprets the query and returns the results. The relevant passages are returned for the query and then ranked before finally returning results to the user. There are a few issues associated with text mining related to ITQ. One of the biggest problems is the ambiguity of the word's meaning, as one word may have many meanings, so the correct results may not be returned for a particular query.

Hanum et al. (2013) have developed a parser system for the Malay translation of the Qur'an. The system checks the correctness of the grammar as per standard Malay Grammar rules, using Earley's algorithm. The Malay translation of the Qur'an, selected for the project, has 40,290 words. Each sentence is parsed by initially splitting each sentence or verse into multiple sentences or parts. For each word in the sentence, the grammatical rule is predicted. In the end, a parse tree showing all the rules is developed for all the sentences. The sentence structure of the Malay translation is verified against Standard grammar rules. The standard Malay grammar is evaluated by performing 42 experiments. It was found that many sentences did not follow the standard Malay grammar rules. The new set of rules consists of 115 rules and the original standard Malay grammar has 94 rules.

3.5 Search systems and Qur'anic NLP

Hammo et al. (2007) designed a search engine using query expansion techniques and tested if the search results can be improved or not. Each verse is tokenized and the words are then stored in three different indexes which are Vowelized-Word Index, Non-vowelized-Word Index, and Root Index in RDBMS. The Vowelized Index holds all the distinct words of the Qur'an without any preprocessing. Hence diacritical marks are present on them. The non-vowelized index is formed by removing diacritics from the words and the root index is finally used to store the root words which are formed by stemming all the words to their base. There is also a Thesaurus in the database which holds the synonyms of Qur'anic words. Three experiments are performed. In the first experiment, the query word was used for search and it was found that the results were not good as the input query may or may not match exactly with the desired word from the Qur'an. In the second experiment, stemming of the query word was done and then it was searched in the root index. Once the word was found in the table, the query expansion was done by finding the related word in vowelized and non-vowelized tables. The experiment showed great improvement in results. In the last experiment, query expansion was done using Thesaurus, which searched the verses for query words based on the synonym approach. This also showed great results. Overall, it was concluded that query expansion is a great way to perform a search on Qur'an.

Al Gharaibeh et al. (2011) have studied the use of formal methods for Qur'anic NLP-based search engines. Formal methods are techniques that are used to describe and specify a software tool. These specifications are expressed using Z-notation. The authors first informally describe the schema, then define the syntax of operations to be

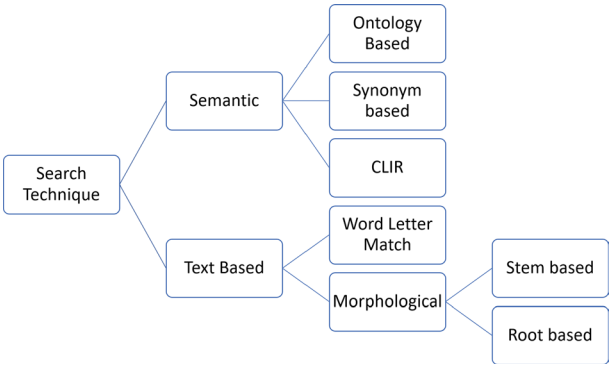


Fig. 12 Types of existing search tools for Qur'an

performed for searching, and finally define axioms that will help in performing operations. There are three types of search techniques included, which are text-based, stem-based, and synonym-based schemas. Schemas are developed for each approach by performing some preprocessing such as stop words and diacritics on the Qur'anic text. In the first approach, the exact query word is matched with the words of the verses. In the stem-based approach, word stemming is performed for query words and words of the verse. In the synonym-based approach, synonyms are used to perform the search. The search engine is developed using Z-notation which has its own syntax. Z/EVES tool is used to develop the schemas and then evaluate the performance of the system developed.

Yunus et al. (2010a) have designed a Qur'anic search framework that performs CLIR with the aid of semantic and stemming analysis on the query. The system takes a query from the user. It is then converted into a target language equivalent query. The system can answer the results in one of the three languages from Arabic, Malay, or English. After the translation, each word of the query is passed through a semantic analysis in which synonyms are provided for each word. A stemmer then converts these words into stems. After this pre-processing, the target verses or documents are retrieved by matching each word from the updated query. It was concluded that the semantic-stemmer-based technique has retrieved far better results as compared to a simple query search. The test was also performed with only a semantic-based search. However, the evaluation showed that recall for the semantic-stemmer-based technique was far better.

Alqahtani and Atwell (2017) have discussed various search techniques and tools for Qur'an. Most of these tools have been discussed in this paper. They evaluated these tools against 13 different parameters to see their drawbacks and limitations. The existing search tools can be classified into two main categories which are text-based searching and semantic searching. Both these categories can be further sub-categorized as shown in Figure 12.

The existing tools apply one of these techniques to answer the queries of users. These tools were evaluated at common criteria which encompass search features, the precision of results, database size, query and answer type, response time, and database contents. The evaluation results can be summarized as follows:

1. Limitation for retrieving all requested information from the Qur'an related to the query.
2. Many tools allow one word-based query and do not allow search based on concepts, phrases, or topics

3. The ontology-based search tools do not cover all the concepts. Only a few ontologies are covered and thus they fail to deliver results against all concepts
4. None of the tools solves the ambiguity problem of retrieved results. By ambiguity we mean that results for words having the same spellings.
5. The tools do not apply modern NLP techniques such as parsing or spell check.
6. These tools do not have Name Entity Recognition especially designed for Qur'anic text.

As discussed in the Ontology-Based Technology section, Khan et al. (2013) proposed a semantic search tool based on an ontological approach. However, the ontology does not cover the complete concepts of the Qur'an. The proposed ontology covers animals and birds in Qur'an. The ontology was developed after an analysis of each verse of the Qur'an. SPARQL query language was used to perform the search. The query is taken in the form of a natural language question in English. The system converts the question into system specific query format. The questions like "Which animal is used for riding?" are answered correctly. It was also recommended to extend the proposed framework further in a proper search tool that can take input through a GUI and performs a search using Qur'anic explanations and Hadith as reference.

Alqahtani and Atwell (2016) proposed a search tool named as Arabic Qur'anic Semantic Search Tool. The tool is based on an ontology of the Qur'an and utilizes both information retrieval and semantic search approach. The query entered by the user undergoes various NLP checks such as stop word removal, stemming, and POS tagging. Semantic tags and synonyms are provided for each word. The search is first performed using the concepts from the ontology. If no results are retrieved then, the search is performed by word matching technique. Finally, the results are provided with the score for their relevance to the query and then ranked before presenting to the user. The tool also deals with the ambiguity of results for words having the same spelling and provides detailed results with concepts of each word. The tool also addresses the issue of the lack of NER for the Qur'an.

Shoaib et al. (2009) developed a semantic search tool to overcome the deficiencies of a simple keyword-based search system. The framework is developed for *Surah Al-Baqarah* of the Qur'an (Chapter 2) initially. To perform the semantic search WordNet is utilized, which provides different synonyms and antonyms of words. We know that a single word can impart multiple meanings, so the query word is allocated a single interpretation at the start of the search. The system then searches for the verses that contain the word and followed by the synonym-based search. The retrieved results are returned in a prioritized order with verses retrieved for exact synonym-based search first and the weak synonym-based results last. 80% more relevant results were achieved as compared to a simple query search.

To overcome the irrelevancy issue and poor retrieval results for keyword-based search and domain-specific ontology in semantic search, Alqahtani and Atwell (2015) designed Qur'anic Semantic Search Tool (QSST). The system aligns existing ontologies to provide better results. Initially, the system performs a concept-based search with the help of ontology, and then it performs a search by word matching technique. This technique achieves the best of both worlds, keyword and semantic search. The results are ranked and redundant verses are eliminated.

Pitchay and Ridzuan (2016) have also reviewed the strengths and weaknesses of existing Qur'anic search tools. They also designed a new ontology by aligning Qur'an ontology and medical domain-specific ontology. The majority of the ontologies developed for the concept-based search for the Qur'an, as discussed earlier, are domain specific. They do

not cover the whole Qur'an. Only a few ontologies cover the whole Qur'an but they do not cover all the concepts and thus fail to retrieve results for the complex query. In this work, an ontology for Qur'an taken from Qur'an ontology.com has been aligned with one of the medical field-specific ontologies. Arabic WordNet is used to perform lexical matching between two ontologies. Finally, the same concepts are matched from one ontology to another using a fuzzy algorithm. The framework has two components, loading, and matching. The former component will make objects for the ontologies and the latter will combine by mapping similar components. The algorithm works by matching some persons from the Qur'an ontology to some human diseases from the medical ontology. The results of the experiment were not great and output accuracy was quite low. The main reason for the failure is that there was inconsistency in the integrated ontologies. The future target is to fix the inconsistency issue and develop an intelligent Qur'anic search system.

The overall summary of the section is presented in Table 8.

3.6 Classifier systems and Qur'anic NLP

Al-Kabi et al. (2005) have developed an automatic text categorization (ATC) tool to classify the verses of the Qur'an. The verses are classified under the various subjects designated by Islamic scholars, such as Faith (Iman), Prayer (Salah), Pilgrimage (Hajj), etc. The tool has been developed using Microsoft Visual Basic. The algorithm is quite simple and easy. Initially desired verse is selected from a *surah*. Then each verse is normalized into features. For each feature, verses are searched from the corpus that contains that feature or word. With the help of all the verses extracted, the relevant subject of the initial verse is allocated. Allocation of the subject is done by calculating the subject percentage for each feature in the verse. The subject with the highest percentage is then tagged to that particular verse. The prototype has been developed for *Surah Al-Fatihah* (Quran Chapter 1) and *Surah Yaseen* (Quran Chapter 36). The accuracy of the results reaches up to 91%. To increase the accuracy further, the corpus is to be increased and once the whole Qur'anic corpus is included in the project, the precision will be quite high.

Ontology developed by Ali and Ahmad (2013) is also helpful in the classification of Qur'anic verses on the bases of themes. Verses belonging to Manners and Faith theme are classified. Sub-themes are available within these classes. This hierarchical approach helps in the classification of Qur'anic verses at various levels. This thematic classification can be helpful for the study and search of various themes discussed in the Qur'an.

Nur and Nurul (2021) have conducted multiple machine learning experiments to find the interrelation between Qur'an and Hadith and the text categorization. Classification algorithms like Naive Bayes, KNN, and SVM were trained on datasets consisting of both Qur'an and Hadith. TF-IDF-based features were used for training the models. The experiments concluded that SVM performed better in the classification of the text and finding interrelated topics between Qur'an and Hadith.

3.7 Topic extraction/categorization and Qur'anic NLP

Alhawarat (2015) have applied LDA for the topic modeling. It is one of the most popular probabilistic methods used in NLP for topic modeling. LDA method was chosen in the previous studies that outperform K-means clustering in most of the experiments. The experiment was performed on the corpus of *Surah Yousaf* (Chapter 12) of the Qur'an. The possible list of

Table 8 Search systems

Sr.	Research	Issue addressed	Technique applied or contribution
1.	Hammo et al. (2007)	Lower accuracy in Qur'an verses search system based on a word	Query expansion using search based on all variants of query word and its synonyms
2.	Al Gharaibeh et al. (2011)	Unambiguity in conventional methods for Qur'an search engine tools	Development of search engine based on text, stem, and synonym searching using Formal method and Z notation
3.	Yunus et al. (2010a)	Low accuracy of results for query-based search	Developed CLIR search tool for Arabic, Malay, and English with the aid of semantic and stemming approach
4.	Alqahtani and Atwell (2017)	Evaluation of drawbacks of existing Qur'anic search tools	Evaluation of search engines based on criteria such as search techniques, ontology, database size, etc
5.	Khan et al. (2013)	Concept based queries for Qur'an	Ontological approach. Animal domain ontology is developed
6.	Alqahtani and Atwell (2016)	Concept-based search, lack of comprehensive ontology, NER for Qur'an	Developed AQSST, which performs ontology-based search aided by keyword search
7.	Shoaib et al. (2009)	Low accuracy of results with simple query	Semantic search tool based on WordNet. Verses are searched for the words in a particular sense
8.	Alqahtani and Atwell (2015)	Low accuracy of keyword search and domain specificity of ontology-based search	Developed QSST search tool that has an alignment of ontologies for better conceptual results. QSST also supports word matching technique for search
9.	Pitchay and Ridzuan (2016)	Review of gaps in existing Qur'anic search tools and development of medical ontology	Aligning Qur'an upper ontology and medical domain ontology using fuzzy logics. The accuracy was low and required fixes in ontology integration

topics was already defined. The probability of a word belonging to a particular topic is defined with the help of the equation given below:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|x_i = j)P(x_i = j) \tag{1}$$

Here w_i is the ith word of a document and, x_i is the i^{th} topic. The experiment was performed on verses based on original words of *surah*, their stem, and their roots. However, it was observed that the results with roots were not acceptable. Various experiments were performed with variations of LDA. The results showed that except for a few, most of the topics included a mixture of more than one topic and did not provide correct context. It was concluded that LDA can be a better method for routine NLP topic modeling but for the Qur'an, the model does not work.

Hassan et al. (2015) also used machine learning for topic categorization but they adopted a supervised learning technique. They performed the text categorization of Malay translation of Qur'anic verses using the KNN algorithm. In the KNN algorithm, the model is trained on data and then test data is allocated a label depending on the similarity with neighbors from training data. For the experiment, vectors of verses were created after preprocessing. The training was done on 800 verses belonging to seven different topics and later test was performed on 200 verses. The recall for seven categories ranged from 0.74 to 0.90. The low recall for the two categories was due to common words involved in both categories. The results were great in the sense that the recall achieved is on the higher side and also that algorithm is quite easy as compared to LDA as discussed earlier. The results can further be improved by increasing the dataset.

Alshammeri et al. (2020) have used the approach of embeddings for the topic modeling of the Qur'an. For the experiment, instead of using word vectors, document vectors were created based on original verses of the Qur'an. The document vector approach proved to be more useful to retrieve the semantically close verses for a given verse. These vectors were then fed as features for a K-Mean clustering algorithm for the topic modeling. All the verses of the Qur'an were categorized into 14 clusters. The experiment showed that the clusters based on document-level vectors helped in achieving semantically rich topics for Qur'an.

In contrast to machine learning methods, Yauri et al. (2012) categorized the topics using ontology. Ontology is the most widely used method to make computers interpret and understand concepts. Since knowledge and topics are usually centered around a particular concept, this can be a useful method for topic categorization. The framework was developed using the already-built Leeds University ontology that only discussed nouns of the Qur'an. This ontology was further extended in this project and concepts of acts were introduced. Ontology was developed using Protégé in which different concepts are linked in a hierarchical fashion where the most general concept is at the top. At the top, there are 15 topics such as living creatures, location, and religion. Under the umbrella of these concepts more specific concepts are defined. Due to this semantically rich categorization, verses belonging to a particular topic can be easily extracted as each verse is marked under some particular concept or topic.

3.8 Knowledge extraction and Qur'anic NLP

Saad et al. (2013) have described a set of rules that can be helpful in the extraction of knowledge from the Qur'an based on an English translation. They have adopted ontology as a source of knowledge extraction for the Qur'an. This ontology is based on three layers,

the meta concept of the Qur'an, the domain ontology layer of Salah, and the last layer that bridges the first and the second layer.

Machine learning methods are widely used in several tasks. Siddiqui et al. (2013) adopted an unsupervised machine learning approach and used LDA to discover the thematic structure of the Qur'an, which will be helpful in knowledge extraction based on some particular theme. They considered one *surah* of the Qur'an as a document and discovered the themes in it. The original Arabic Qur'an was used for the experiment. After the initial NLP preprocessing, data was reduced to 24 *surahs* with 417 terms only out of 114 *surahs*. The reason for doing this reduction was to remove the less frequent terms that can hamper the results. As described in the previous section, LDA works by allocating topics to different words in a document. The topic with the highest probability is assigned to the document. The experiment results showed that the model correctly classified the Meccan and Medinan Chapters of the Qur'an based on the difference in words. The results are promising and proved that LDA can be applied for topic-based knowledge extraction from the Qur'an.

Mohamed et al. (2015) used the supervised machine learning approach of a decision tree classifier to find the answers to questions from the Qur'an. The model was devised for the original Arabic Qur'an. The system was backed by a Qur'an ontology developed by the integration of two different ontologies. The ontology covered 1200 concepts. The framework converts questions and their possible answers to vectors and then finds the semantic relativity of the answers' vectors to the question vector using a decision tree classifier. A keyword matcher is also used to support the results. These vectors are then classified into direct, related, and irrelevant classes. The training was performed on Fatwa questions taken from a website as these are the usual questions asked by the people. The testing results showed that accuracy reaches up to 74.53% The results can be further enhanced if only direct or non-relevant answers are chosen, as the relevant class is generic and it can have answers from both categories.

Ontology can also be used for knowledge extraction. As we have seen earlier, ontology is a conceptual dictionary that categorizes different concepts under a hierarchy. Yauri et al. (2012) have adopted the Web Ontology Language (OWL) to perform a conceptual search on the verses. They reused the ontology developed by Leeds University but introduced more concepts in it related to various acts and worships in Islam. Although it does not cover all the concepts discussed in the Qur'an, more complex queries can be performed with the help of a description logic query system as compared to other ontologies developed. The Manchester OWL syntax is used to perform the query. The system extracts semantically rich answers for the queries as the data is linked in an inheritance-concept system. Although, complex queries can be performed, but the system lacks a user-friendly interface where a query could be entered in a natural language way. However, the team focuses to cover this issue in the future.

The ontology approach is also adopted by Ta'a et al. (2017) but it is more user-friendly. Instead of performing some syntax-based queries, a graphical user interface allows the user to select a particular topic or theme and two sub-themes from drop-down menus. After the selection of the themes and sub-themes, a query can be performed by entering a keyword that retrieves the relevant verses based on the pattern matching technique. The results achieved up to 90.4% accuracy and were also authenticated by scholars. The system provides great results, but it also covers only three themes that are related to Allah, Angels, and Unseen. Ahmed and Atwell (2016) worked on improving the number of topics and concepts covered by an ontology for knowledge extraction. Their ontology includes the merger of previously existing two ontologies, an ontology extracted from the Qur'an and

Table 9 Knowledge extraction techniques

Sr.	Research	Methodology	Technique applied	Outcome or contribution
1.	Siddiqui et al. (2013)	ML	Unsupervised learning-LDA	Correct clustering of Meccan and Medinan Chapters of the Qur'an
2.	Mohamed et al. (2015)	ML	Supervised Learning- Decision tree	Bayyan: a tool that performs query search and returns answers based on semantic similarity. Accuracy: 74.53%
3.	Yauri et al. (2012)	Ontology	OWL-based knowledge extraction	Semantic answers extraction using multi-theme levels, syntax-based query, not user friendly, limited concepts
4.	Ta'a et al. (2017)	Ontology	OWL-based knowledge extraction	Semantic answers extraction using multiple themes, keyword-based query, user friendly GUI, limited concepts, Accuracy up to 90.4%
5.	Ahmed and Atwell (2016)	Ontology	Merger of multiple ontologies	Conceptually rich answers extraction, Accuracy: up to 85%
6.	Saad et al. (2013)	Ontology	Use of set rules and layers	Knowledge extraction related to Salah

Tafsir and an ontology extracted from the reviews written by scholars on various verses. The information from these three types of ontologies is merged to form a new broader ontology that covers more concepts and can answer more queries. The system was able to produce up to 82% accuracy. The results were further improved up to 85% by using a PROMPT tool, which helps in merging small and medium-sized ontologies.

Both Machine Learning based and Ontology based methods can be applied to extract knowledge from the Qur'an. Both these methods are under continuous research and improvement. Various techniques discussed in this subsection for knowledge extraction can be summarized in Table 9.

3.9 Speech processing and Qur'anic NLP

All Muslims around the world either Arab or Non-Arab recite the Qur'an. The Qur'an needs to be recited properly according to the rules specified in the science of Tajweed. Moreover, many Muslims memorize the complete Qur'an by heart. Therefore, it is the need of the hour to empower Muslims with state-of-the-art speech-based learning solutions for the recitation of the Qur'an. The foundation of this would be a speaker independent automatic speech recognition system for the complete Qur'an. Much work has been done in this area and some of the important work is explained in this section.

Brierley et al. (2014) have worked on a set of consonants in the Qur'an that provide the prosodic effect of *Qalqalah* (vibration). Prosody is the pattern of rhythm that how the voice of the speaker rises and falls while speaking. The tool takes the Qur'an as input and using NLP techniques returns the words in the verses which contain *Qalqalah* consonants. A tool named Semantic Pathway is used to find out the types of words having *Qalqalah* effect. It was found out that the most frequent *Qalqalah* words used in Meccan and Medinan *surah* are different. Another contribution is the development of a pronunciation guide for the *Qalqalah* words, which will be helpful for the recitation of the Qur'an.

Ibrahim et al. (2008) provide a review of techniques used in different steps of speech recognition of Qur'an. There are four major stages in speech recognition and these steps are pre-processing, feature extraction, training and testing. Pre-processing is done to improve the readability of audio data. The techniques used in pre-processing are Endpoint Detection, Noise filtering, Smoothing, and Channel Normalization. For features extraction techniques like Linear Predictive Coding, Perceptual Linear Prediction, Mel-Frequency Cepstral Coefficient (MFCC) and Spectrographic Analysis are used. And for Training and Testing purposes Hidden Markov Model (HMM), Artificial Neural Network (ANN) and Vector Quantization (VQ) are used. The authors concluded their research by noting that for feature extraction the most suitable method is to use MFCCs, while for training, HMM is most suitable.

Ahmed and Abdo (2017) also explained the same techniques as explained by Ibrahim et al. (2008) and also proposed a technique for Qur'anic verses verification. The audio recitations from the verified scholars are stored in the database. For testing any new recitation the audio signal is aligned with the Qur'an text after the preprocessing steps and this is the most challenging part. The MFCC feature vectors of this recitation to be tested is compared with the verified recitations and differences in recitation are notified to the reciter.

Qayyum et al. (2018) used a deep learning approach to address a less solved area of speech recognition. They applied Bidirectional Long Short-Term Memory (BLSTM) model to identify the Qur'an reciter. The model takes recitations of various lengths from five different reciters and extracts the MFCC features from them to feed as an input

to BLSTM. The results achieved are exceptional as accuracy reaches up to 99.89% for identifying the correct reciter for 3 seconds of recitation. These results are far better than other powerful models like SVM which remained below 90%. The model also surpassed the conventional ANN model that could achieve an accuracy of 91.8%.

Qur'an is a sacred book and it must be recited with utmost care as by a minor change in the pronunciation the meaning of that word changes. Mohammed et al. (2015) proposed a four-staged method for verification of Qur'an recitation using speech recognition techniques. These stages are input preparation, feature extraction by MFCC, audio match and training followed by testing of the model using HMM. The matching stage is significant as it will search the MFCC feature of input recitation in the database of already stored authentic recitations MFCC features. If the features match then recitation will be considered authentic.

Jamaliah Ibrahim et al. (2013) investigated the Tajweed checking system using speech recognition to complement the conventional methods of Qur'anic learning with the latest technology. The authors provide insights into the use of MFCC for feature extraction and HMM method for classification part of speech recognition. The initial system was tested for *Surah Al Fatiha* (Chapter 1), which results in a recognition rate of 91.95% on verse level and 86.41% phoneme level. These results show that the developed system has the potential to be adopted in the education system to support the traditional methods in Qur'anic learning process.

Muhammad et al. (2012) developed an intelligent system E-Hafiz that helps in recitation and memorization of Qur'an. Recitation. The availability of Qur'an experts is not common in non-muslim countries which hinders the process of learning and memorizing Qur'an. To develop E-Hafiz, MFCC feature extraction is adopted. MFCC features of reciter are compared with that of experts. Reciters recite verses from the Qur'an in the presence of an expert who points out the mistakes. Expert rectifications are compared with the mistakes pointed out by E-Hafiz system to measure the accuracy. The accuracy of this system on a group of men, children, and women is 92%, 90%, and 86%, respectively.

Amrani et al. (2016) use simplified phonemes in Automatic Speech Recognition (ASR) for Qur'anic recitation. CMUsphinx toolkit in used in the project. Data used is the recitations of the first and last 3 *surahs* of the Qur'an. The audio data is of almost 40 minutes from male speakers. Different training configurations are tried and the best result was obtained at 32 dimensions of the Gaussian mixtures. It was concluded that by using a simple phoneme list instead of romanized phonemes, which are difficult to generate for the whole Qur'an, it is possible to build an ASR for the complete Qur'an by adding more recitations of the renowned reciters to the database.

Abro et al. (2012) tried to automate the Qur'an memorization process using speech recognition techniques. For feature extraction, MFCC is used and ANN is used for acoustic modeling and pattern recognition. The data set used has twenty utterances of the last *surah* of the Qur'an (Chapter 114), recited by a fluent reciter of the Qur'an. A dataset for recitations with errors is generated artificially by removing some words from original recitations. Classification of correct verses with errors is done. Experiments failed to differentiate between the correct recitations and recitations with errors and give false positives. It was concluded that this simple technique is not suitable for automating the Qur'an memorization, And further work is needed to be done.

Yekache et al. (2012) discussed the initial steps towards making the Qur'an reader controlled by speech commands. In this regard, they gathered speech data for names of all 114 *surah* of the Qur'an, some reciter's names, and some other commands. They adopted

HMMs for acoustic model training and used the CMUSphinx toolkit for developing this system.

Tabbal et al. (2006) presented the delimitation of Qur'anic verses using speech recognition techniques. The amount of Qur'anic recitations available online for free is promising but these do not have verse by verse trimmed. So, there is a need to make an automatic system that trims the recitations of the Qur'an on verse level. The toolkit used in this work is CMUsphinx. The data set is the recitations of *Surah Al-Ikhlas* (Chapter 112) by professional reciters and by normal reciters who do not follow the tajweed rules strictly. The mean recognition ratio on 5 professional reciters who recited in the Tajweed style is 90% and 8 professional reciters who recited in Tarteel style is 92%. Experimental results on the recitation of normal Arabic speaking 20 males is 90% and 20 females are 85%.

Al-Bakeri and Basuhail (2017) built an ASR for tajweed checking and integrated it into a self-learning environment that must ensure the Qur'an is pronounced correctly as per the tajweed rules. For feature extraction, MFCC is used. HMM is used for acoustic model training and Gaussian mixture density is used to calculate the state emission probabilities. *Surah Al-Ikhlas* and *Surah Al-Rahman* (Chapter 55) recited by ten reciters is used as input. Two scenarios are followed, one is the use of phonemes and the second is the use of syllables. The use of syllables performs well when the data is small and gives 100% accuracy for *Surah Al-Ikhlas*. In results of using phoneme accuracy rate of 89.47% is achieved.

Putra et al. (2012) tried to minimize the difficulty in the learning process of the Qur'an using speech recognition and integrated the system with a learning software. For speech recognition, MFCC features and Gaussian Mixture Model (GMM) modeling are used. Interactive multimedia software is built in a prototype stage. Accuracy for pronunciation of reciters is 70%, 90% for recitation law, and 60% for a combination of these two. The results show that performance is poor as the dataset used is small.

Tabbaa and Soudan (2015) worked on computer-aided training for learning the Qur'an recitation. HMM-based ASR initially recognizes the recitation phones. Only two classifiers are trained, one for differentiation between emphasized and non-emphasized pronunciation of letter R and the second classifier to discriminate closely related letters. Data set used is the telephone calls on a TV program where a professional reciter recites a page from the Qur'an and then listens to the callers'. The ASR system gives 97.6% word level accuracy. For phone-level classification, four different classification algorithms are tested and the best one is chosen for each scenario. These four algorithms are SVM, Neural Network Multi-layer Perceptron (MLP), Bagging, and Random Committee. To test the performance of the complete system 60 minutes of audio data from 18 female and 14 male reciters are used. The previously mentioned 97.6% accuracy is reduced to 84% if mispronounced phones are considered a word-level error. After integrating the classifier, the accuracy is improved to 91.2% as it reduced the phone level of false positives and false negatives.

Satori et al. (2007) studied some fundamentals of Arabic speech recognition using the CMUSphinx toolkit. The data set was made in-house by 6 Moroccan male speakers, each of them spoke ten Arabic digits. Each digit is repeated 5 times by each speaker and hence a total of 300 utterances are there in the data. The mean recognition ratio of the testing experiment on the three male-person utterances of these 10 digits is 86.66%, 86.66%, and 83.33% respectively. Hello_Arabic_Digit application is presented so that it would be adopted in the Arabic speech recognition system.

El Amrani et al. (2016) investigate the use of simplified Arabic phonemes in building a phonetic dictionary for a speech recognition system. Normally Romanized phonemes are used in phonetic dictionaries required to train ASR. The audio data used are the recitations of famous reciters of the first and last three *surahs* of the Qur'an. CMUsphinx toolkit is used and the building of a phonetic dictionary consisting of simplified Arabic phonemes from the Qur'anic text is automated through a computer program. Hence, the transcription file, phonetic dictionary, and list of phones are automatically generated by the developed program. Experiments have been done with different training settings and the lowest word error rate obtained is 50.0% and 55.7% while using 90% and 80% of the audio data in training respectively.

Ahsiah et al. (2013) propose a tajweed checking system to support the learning of the Qur'anic recitation. The proposed system listens to the reciters' recitation and matches it with the recitations of the experts and finds out the differences between them and hence tells the mistakes to the reciter. The proposed system used MFCC for feature extraction and HMM for feature classification.

Mohammed et al. (2017) investigate the phoneme duration at *Madd*, *Ghunna* and some other letter characters. These are a few rules from the important Tajweed rules that need to be followed while reciting the Qur'an. To calculate the duration of *Madd* and *Ghunna* in recitation data is collected from the recitations of expert reciters. This system is used to point out any mistakes in *Madd* and *Ghunna* rules of any person's recitations. 600 words of data are collected by recitations of 10 Reciters. And from the recitations mean duration of each the of the *Madd* and *Ghunna* is calculated. The subsection is summarized in Table 10.

3.10 Quranic corpus and Qur'anic NLP

Alrabiah et al. (2014) have contributed to classic Arabic NLP by creating a large classical Arabic corpus of more than fifty and a half million words. The corpus consists of six domains and the major portion is dominated by religion with 46.73%. They have also performed two empirical studies. Both these studies are dedicated to finding the collocations in classical Arabic. For this purpose, they have performed eight statistical association measures. For the first study, QAC is chosen. This first study is further broken into two phases. In the first phase, mean average precision (MAP) is measured for the association rules by making the least frequent part of the collocation node while in the second part the most frequent part is made the node. For the Qur'anic corpus, collocation extraction with the least frequent word as node gave better results. For choosing the most frequent word as a node, the MAP for all association results drops significantly, because the more frequent part will have more association with other words. There is a great drop in MAP for log-likelihood, although it shows good performance in the first phase. For the second empirical study, association measures were checked on the newly developed large corpus that is KSUCCA. The experiment aimed at investigating whether the association measures change by changing the size of the corpus or not. It was observed that scores for all the measures dropped quite significantly. However, it was observed that MI.log-frequency and log Dice are the most suitable in terms of their MAP for both small and large corpora. The results of the eight association rules and their comparison are shown in Table 11:

Table 10 Speech recognition work for Qur'an

Sr.	Research	Issue addressed	Technique applied or contribution
1.	Brierley et al. (2014)	prosodic effect of vibration of words for Qur'an	Investigation of various vibration words in Qur'an and pronunciation guideline development
2.	Ibrahim et al. (2008)	Review of techniques for speech recognition	Concluded that MFCC and HMM are the best method for feature extraction and training respectively
3.	Ahmed and Abdo (2017)	Review of techniques for speech recognition	Description of various methods involved in speech recognition for Qur'an recitation
4.	Qayyum et al. (2018)	Identification of the Qur'an reciter	Used deep learning for speech recognition and identification with 99.89% accuracy
5.	Mohammed et al. (2015)	Integrity and authenticity of the Qur'an in recitation on social media	Speech recognition methodology proposed based on MFCC and HMM
6.	Jamaliah Ibrahim et al. (2013)	Tajweed checking	Tajweed checking with speech recognition based on MFCC and HMM
7.	Muhammad et al. (2012)	Qur'an memorization with correct Tajweed	Speech recognition based system with accuracy above 86%
8.	Amrani et al. (2016)	Correct Qur'an recitation identification	Use of simplified phonemes in Speech recognition
9.	Abro et al. (2012)	Qur'an memorization	Neural network based speech recognition using MFCC
10.	Tabbal et al. (2006)	Delimitation of verses using speech recognition	Speech recognition of Verses from various reciters using CMUsphinx tool
11.	Al-Bakeri and Basuhail (2017)	Correct Tajweed of the Qur'an	ASR based on MFCC and HMM using phonemes and syllables
12.	Putra et al. (2012)	Qur'an recitation learning	ASR based on MFCC and GMM. Prototype gives up to 60% accuracy combined for pronunciation and recitation laws
13.	Tabbaa and Soudan (2015)	Qur'an recitation learning	HMM-based ASR for classification of errors in recitation using different AI algorithms
14.	Satori et al. (2007)	Arabic speech recognition	Development of "Hello Arabic" App with accuracy of 83.33% for Arabic ASR

Table 10 (continued)

Sr.	Research	Issue addressed	Technique applied or contribution
15.	El Amrani et al. (2016)	Investigation use of simplified Arabic phonemes for speech recognition	Use of the Qur'anic phonemes with CMUsphinic toolkit
16.	Ahsiah et al. (2013)	Correct Tajweed checking	ASR based on MFCC and HMM for the identification of mistakes during recitation
17.	Mohammed et al. (2017)	Mistakes identification during particular letter recitations of the Qur'an	Mean time of recitation calculation of "Madd" and "Ghunna" to estimate the correct duration of their recitation

Table 11 Performance of algorithms against corpus

Measure	MAP with	
	Qur'anic corpus (%)	KSUCCA (%)
MI.log-freq	76.68	37.02
Log dice	72.64	35.12
MI	61.44	5.19
T-score	68.70	14.92
Raw frequency	67.26	13.39
Minimum sensitivity	72.77	31.72
Log-likelihood	76.66	26.10
MI3	77.01	34.12

3.11 Q/A systems and Qur'anic NLP

Hamed and Ab Aziz (2016) developed a question-answering system using an artificial neural network to classify the verses related to pilgrimage and fasting. The dataset used for this tool is the English translation of the second chapter of the Qur'an. The tool consists of three components. The first component is the question analysis module, which performs preprocessing on the user query to clean it and then performs the expansion using WordNet and Islamic terms. The document retrieval module is the second component, which is based on a neural network. Model is trained in this component at 150 verses belonging to three categories of fasting, pilgrimage, and none of these. The final component is the answer selection module. With the help of the first module where the question was converted into a machine-interpretable form, the relevant verses are retrieved using the N-gram matching technique. The relevant verses are ranked where the verse with the most matched words with the question is placed at the top. The model has provided great results. The F-score of verses classification is 90% and the answers retrieved by the tool achieved the F-score of nearly 87%.

Al-Bayan, the question-answer system developed by Mohamed et al. (2015) uses a supervised machine learning approach and keyword matching. The system provides the answer by a complex system, which makes it sure that the answer returned is semantically related to the question. The system makes a semantic relatedness check by calculating cosine similarity between question and answer vectors. In the second stage, it calculates the keyword matching score. These two scores are used as input to the decision tree that calculates if the answer falls in the direct, related, or irrelevant category. The tool was evaluated on common Fatwa questions and answers.

Hamoud and Atwell (2017) have contributed by developing a corpus that consists of questions and answers related to knowledge of the Qur'an. The corpus contains 1000 questions and answers in Arabic and 500 in the English language. To make a corpus or dataset related to the Qur'an requires extra care about the authentication of information and knowledge as it is a divine book according to Islamic belief. Four sources were used to develop the corpus. A large set of questions and answers was retrieved from different websites. This data is important as it is closer to the people asking various questions related to Islamic concepts. Some questions are retrieved from the Qur'an itself. The third source is the questions asked by the Muslims by scholars of the Holy Mosque in Makkah and the fourth source is from a survey. All the data collected is merged and unified and is made useful

by cleaning it. This huge effort can prove very useful in developing tools and platforms that provide authentic and relevant answers to people efficiently. The corpus has not been made public yet as the work continues to enhance the dataset by creating variants of current questions.

Recently, a Qur'anic Reading Comprehension Dataset (QRCD), which is related to the automatic extractive Question Answering system competition announced for the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT), 2022 for Language Resources and Evaluation Conference (LREC). Details of the dataset, competition, and evaluation criteria are described in (Malhas et al. 2022). Malhas and Elsayed (2022) extended the earlier work by introducing CLassical-AraBERT, a new AraBERT-based pre-trained model which is further pre-trained on approximately 1B words classical Arabic (CA) dataset. It is intended to complement MSA resources used in pre-training the initial model. And finally, they leveraged cross-lingual transfer learning from MSA to CA. Moreover, the authors introduced a new metric, pAP (partial average precision). The new metric integrates partial matching in the evaluation over multi-answer and single-answer MSA questions.

3.12 Interface systems and Qur'anic NLP

An interface system is used where a user can easily interact and can get the answers to queries without having any knowledge of programming. Interfaces can be online applications that are created for the common man. One such system has been designed by ElSayed (2015) for people who are not database experts. The system takes input from users in the form of an Arabic language query. Most of the systems have knowledge stored in their databases in the form of structured data that is returned upon the query. This query has to be in the specific syntax of the database system. The designed system allows the user to input the question in the Arabic language and then converts that question into a specific query and returns the output. The system can be divided into two parts. The first part is NLP-based. After the user inputs his query using a graphical user, there is a lexical analyzer scans the whole sentence and performs a spelling check too. The Arabic query is then converted into its English translation. A parser converts the whole sentence into structured parts to make it easy for the computer to understand. After this translation, the second part comes into play. SQL generator converts the English sentence to a resembling SQL query. Finally, this query is applied to the database and the results are returned to the user. The database consists of Qur'anic chapters, verses, words, and *Juz*. There are two types of query modes, imperative mode, and question mode. The former starts with an imperative verb and the latter is an interrogative style questioning mode. The system currently retrieves answers to simpler queries like "Which *surah* contains the word Allah?"

3.13 Authentication/verification and Qur'anic NLP

The Internet is the source of information in the 21st century. There is a large amount of data that is available on the internet today, which keeps increasing exponentially. However, the information contained in this data can be correct as well as incorrect. So, one has to be very careful while browsing and searching such data. Anyone can post anything he wishes and that can be read by anyone. The same is the case with Islamic information. Many people write and discuss Islamic issues online. In these issues, it has been observed that when Qur'anic verses or Hadith are quoted as references, many times the quoted references are

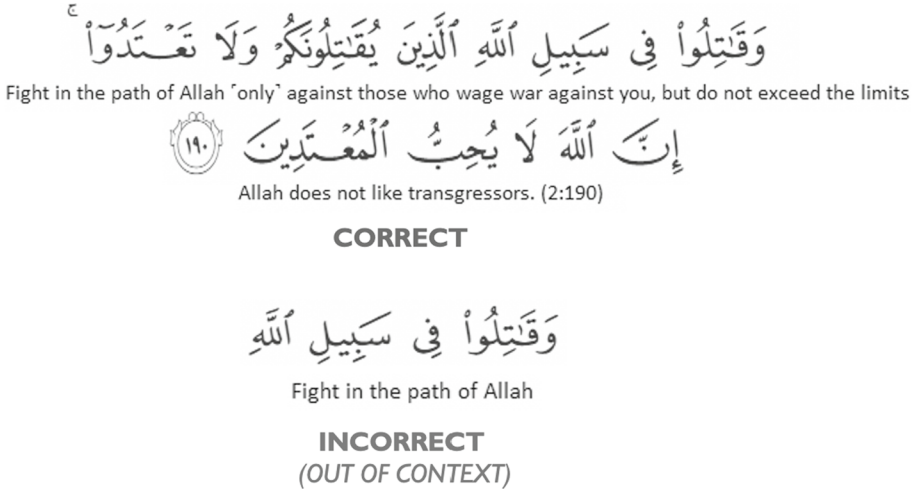


Fig. 13 Correct and incorrect quotation of Qur'anic Verse 2:190

either incomplete, incorrect or totally out of context. This leads to complete misinformation. Special care is required when it comes to referring to the Qur'anic verses. As Qur'an is a sacred book, no change is expected for even a single letter in the whole book. Muslims around the world try their best to make sure that there is no mistake while quoting Qur'anic verses online. However, incorrect quotations can arise due to multiple reasons. Sometimes a word or letter may be missed mistakenly while typing or copying and sometimes incomplete verses are quoted because of Islamophobia. For example, it has been observed that, after the war on terrorism began, a Qur'anic verse was quoted out of context by many people, portraying that Qur'an persuades all Muslims to fight non-Muslims in the name of Allah. The half-quoted verse with its full version is shown in Fig. 13. There is a need for authentication tools that can verify the online available Qur'anic quotes to provide the true knowledge of the Qur'an to everyone.

Alshareef and El Saddik (2012) have designed a tool to verify the authenticity of Qur'anic verses quoted online. Qur'an is either quoted in different issues being discussed online or there are online platforms that provide utility to read Qur'an online. All these Qur'anic texts must be correct. The authentication checker model consists of two parts, which are Qur'anic quote filter and Verifier. In the filter, the quote to be verified is inserted into the system and then diacritics and symbols are removed from it. After that, the verifier part checks if the quoted verse is in the Qur'an or not. If the word-by-word verse is matched then it is marked as correct. If it is a partial quote or if two verses are quoted without a full stop, they are considered incorrect because both these cases change the meaning. In case, no verse matches exactly, the system checks for the nearest possible verses in Qur'an and returns them. If there is no verse similar to the quoted one, then the input quote is marked as incorrect. The system checks for similar verses and exact matches using SQL query. The tool is very useful in terms of checking the authenticity of verses. The authors have reported that they plan to integrate the tool with the online web so that it can be accessed by Internet users to verify the quotes that they are reading.

Alsmadi and Zarour (2017) have also developed an online tool that can help in checking the authentication of Qur'anic verses available online. The tool consists of two parts. The

first part is used for information retrieval and the other part is used to check the authentication of the retrieved verses against a standard relevant verse. The information retrieval part consists of a web crawler. Web crawler works by finding the queried text at online websites and then returning the relevant pages. In this tool, a correct verse is given as a query and the crawler then retrieves the top ten pages returned from the Google search engine. These pages include both Islamic and non-Islamic pages. The verses provided by these pages are then checked with the verifier system. The verifier uses a distinct method that is of using a hashing algorithm to verify the retrieved verse. MD5 algorithm was chosen because of its popularity. It was also verified if font size and color affect the code for the same input or not. After all the checks, the MD5 algorithm was applied. The algorithm is great in the sense that it returns a completely different hashing code even on a change of a single character or diacritic. The experiment was performed on one of the websites and many errors were found in different verses.

3.14 Crowdsourcing and Qur'anic NLP

Crowdsourcing is one of the most popular methods adopted worldwide for NLP tasks. Datasets are built with the help of the public. In this way, a large number of inputs are received and there is variety in data too. This particular method is also adopted for creating NLP datasets, tagging, and annotations tasks. Crowdsourcing help achieve the goal in a much more efficient way. A similar sort of approach can also be adopted for Qur'anic NLP. The contribution that can be achieved by crowdsourcing can include word tagging, similarity checks, annotations, POS tagging, and many others. However, crowdsourcing for Qur'anic NLP can differ from other NLP tasks because a lot of care is required in this case as Qur'an is a religious book of Islam and any wrong input can lead to false information.

Zaghouani and Dukes (2014) have experimented to see if crowdsourcing can be beneficial for Qur'anic NLP or not. They used Amazon Mechanical Turk (AMT) to perform annotations for Qur'an. There are many crowdsourcing platforms available, but AMT is the most widely used for such tasks. There are many workers on the platform who get paid for the problem they solve. Keeping in view that Arabic is not an easy language, the questions were asked with multiple choices, so that the efficiency may be improved. Two different tasks were chosen, POS tagging and grammatical case endings. For the grammatical case endings, 100 words were chosen from chapter 23, and for POS tagging, 200 words were chosen from the same chapter. Since such tasks are not easy for non-Arabs and even some of the tasks are even difficult for expert Arabic linguists, therefore initially a screening test was also taken and only high performers were allowed to perform the tasks. Out of 137 workers, only 24 could pass the test and among them, 17 workers could perform annotation. This low number of workers doing the annotation shows that Arabic NLP, and especially Qur'anic NLP, is not an easy task. The accuracy for grammatical case ending was 50% and for POS tagging it was 63.9%. The results were benchmarked with QAC. The results proved that crowdsourcing is not a good option for Qur'anic or Arabic NLP. It can be concluded with this experiment that crowdsourcing can not be directly applied in this case and tasks related to Qur'an can only be done with the help of experts.

4 Qur'anic NLP tools & resources

In this section, we will highlight the tools that have been used in various Qur'anic NLP tasks. There are many tools and platforms that have been utilized. The tools covered here are those that have directly contributed to Qur'anic NLP.

4.1 QurSim corpus

Sharaf and Atwell (2012), is a large Qur'anic corpus with similar and related verses linked together. This corpus is useful for NLP tasks such as automatic similarity and relatedness detection in short texts as well as machine translation and paraphrase analysis. The corpus includes more than 7600 pairs of related verses collected from multiple scholarly sources. The dataset was incorporated with query pages online to allow the visualization of a given verse and its network of related verses.

4.2 Quranic proposition bank

To address the challenges of Natural Language Understanding, Palmer et al. (2005) created the English Propbank which is a corpus annotated with verbs and their arguments. This corpus was used in the automatic Semantic role labeling (SRL). Later on, Zaghouani et al. (2012); Palmer et al. (2008) created the first Arabic Propbank which was followed by a Qur'anic Propbank (Zaghouani et al. 2012). The Qur'anic Propbank was based on the Qur'anic Arabic Dependency Treebank (QATB) (Dukes and Buckwalter 2010).

The Qur'anic Arabic PropBank (QAPB) is a unique resource as it increased the coverage of the Arabic Propbank by adding the Qur'anic Arabic variety and the semantic usage of classical Arabic religious text and poetic literary Arabic. to create their corpus, the authors used an Arabic root meaning tool as a reference tool to identify the multiple possible meaning of the verbs in the Qur'an.

4.3 Annotation tool

BAMA is an annotation tool developed in 2004 by the Linguistic Data Consortium, the trustees of the University of Pennsylvania. BAMA performs annotation of Arabic text by tagging POS (Buckwalter 2004). This tool has been used in the automatic morphological annotation phase of the Qur'an by Dukes and Habash (2010) with few modifications. BAMA performs the tokenization of the text and the segmentation of words into prefix, stems, and suffix. The tool then analyzes these segments and allots POS. The tool also helps in analyzing noun case endings, verb moods, vowels, and diacritics. BAMA can be helpful in grammatical tasks related to the Qur'an but care must be taken as this tool has been developed for MSA and not for the Qur'an in particular. Many words of the classical Arabic Qur'an are not available in BAMA (Dukes and Habash 2010).

Muhammad (2012) presented a methodology to annotate conceptual co-reference and text Mining the Qur'an. The raw Arabic text of the Qur'an was divided into several morphological units using the Qur'anic Arabic Corpus (QAC).

Later on, Information Retrieval (IR) techniques were used to convert and index the Qur'anic terms. In total, approximately 24,000 pronouns were annotated with their

references. This list of referents was organized into more than 1,000 ontological concepts. This data set is helpful in particular for NLP tasks such as the automatic co-reference resolution.

4.4 Querying tools

SPARQL is a query language used to perform queries on RDF (Resource Description Framework) (spa 2008). RDF is a database for the semantic web, which is an extension of the world wide web. RDF stores the structured and semi-structured data of websites so that it can be queried for useful purposes. SPARQL is similar to SQL language SPARQL has been used by Sherif and Ngonga (2015) where they have stored Qur'an data in RDF format. The purpose of storing data in RDF is to make it usable for many online platforms. Khan et al. (2013) have also used SPARQL query language for the semantic search of the Qur'an in the English language. SQL is a structured query language that is used to perform queries on database (sql). Unlike, SPARQL, it is not designed for the semantic web. SQL can be used to retrieve from and store to database. SQL has been used in many search applications developed for the Qur'an. Qur'any Explorer is a comprehensive tool available online that covers the various concepts and themes found in the Qur'an created by Noorhan Abbas and Eric Atwell.[7]

The corpus of Qur'any Explorer is linked to an ontology taken from Mushaf Al Tajweed which is considered an expert source. The ontology covers approximately 1200 Qur'anic concepts. Scholars can use the Qur'any ontology browser to find a given concept and display the verses related to the concept selected accurately. Qur'any was implemented by using several technologies such as Google Python AppEngine, the NLTK Python Natural Language ToolKit, XML and AJAX JavaScript, and Yahoo! User Interface Library.

4.5 Tools for ontology

Protégé-OWL is an open source editor for Protégé framework, which is used to develop and edit ontology systematically (pro 2004). OWL is capable of exploiting features of Protégé such as visualization, storage formats and user interface (Rubin et al. 2005). Protégé-OWL itself is developed in Java. This technology has been utilized by Ta'a et al. (2017) in the development of an ontology for the Qur'an. Since it is an open-source tool, it can help develop a conceptual search system based on ontology.

Noy and Musen (2003) developed a set of tools for managing multiple ontologies. IPROMPT tool is one of these tools used for merging two ontologies. The tool provides guidelines to the user in each step to make the correct merger. It also provides suggestions to avoid inconsistencies and errors. The other tool is ANCHORPROMPT, which finds correlations between concepts described in ontologies to be merged. There have been various ontologies developed for the Qur'an but most of them are partial. Ahmed and Atwell (2016) merged the partial ontologies using PROMPT tools and achieved a single consistent ontology with a broader range of concepts.

[7] The Qur'any Ontology browser is accessible online at: http://quranytopics.appspot.com.

Table 12 List of tools used for Qur'anic NLP tasks

Sr.	Tool	Purpose
1.	AMT (amt)	For crowdsource work
2.	BAMA (Buckwalter 2004)	Morphological annotation of Qur'anic words
3.	CMUsphinx (cmu)	Speech recognition applications of the Qur'an
4.	PROMPT tool (Noy and Musen 2003)	Ontology merging tool
5.	Protégé-OWL (pro 2004)	Tool for Ontology development
6.	QAC (Dukes 2009–2017)	For annotation rich corpus of Qur'an
7.	SPARQL (spa 2008)	For performing queries
8.	SQL (sql)	For search applications
9.	Tanzil (Zarrabi-Zadeh 2007–2021)	For Qur'anic Translations

4.6 Speech recognition tool

CMUsphinx is an open-source toolkit developed by Carnegie Mellon University. It is a leading speech recognition toolkit adopted in various speech applications (cmu). It consists of four components used for speech recognition and training acoustic models. The tool contains HMM inside it for speech recognition. CMUSphinx has been used in many speech recognition applications that have been developed for the Qur'an. Most of these applications are focused on correct *Tajweed*.

4.7 Crowdsourcing tool

AMT is a crowdsourcing platform by Amazon that helps individuals and companies to outsource their work (amt). AMT is mostly used for survey tasks in which people are paid for their contributions. This helps in having data collected from different sources in a short period. Crowdsourcing can also be utilized for developing datasets for Qur'an NLP tasks. Zaghouani and Dukes (2014) used AMT to see if it is beneficial to use this crowdsourcing platform for Qur'anic NLP tasks. However, the results were not satisfactory as the task was related to the complex domain of grammar that requires expertise. However, AMT can be used in tasks related to Qur'anic NLP that do not require expert skills, e.g., collection of recitations from people, survey related to memorization of the Qur'an, or some other generic Qur'an-related survey.

4.8 Qur'an translation resource

Tanzil is an open source platform that not only provides Qur'anic Arabic Unicode for various applications but also provides translations in many languages (Zarrabi-Zadeh 2007–2021). There are multiple translations available in each language provided by

Fig. 14 Timeline with respect to major milestones based on Qura'nic NLP survey

different translators. The translation can be downloaded and used in Qur'anic NLP applications. The translations can be downloaded in different formats.

Another valuable repository dedicated to application developers is the Qur'andatabase.org[8] as they can easily download 104 translations of the Qur'an in multiple formats such as plain text, CSV, XML, and MSQL among other formats.

4.9 Qur'an corpus resource

QAC has been discussed as a contribution in this paper but it has also been used as a resource in Bentrcia et al. (2018) and Sherif and Ngonga (2015). QAC is a rich corpus that can be especially used in tasks where annotated Qur'anic data is required. QAC provides word-level grammar, syntax, and morphology (Dukes 2009–2017). The corpus is significant as the annotations have been manually verified by volunteers and experts.

Table 12 summarizes the tools used for Qur'anic NLP tasks.

The timeline for milestones related to Qur'anic NLP are shown in Fig. 14. This timeline is as per the survey conducted in this paper.

5 Caveats and potential pitfalls in Qur'anic NLP research

The matter of using technology to obtain an understanding of religious texts is a sensitive one. Languages are well-known to be ambiguous and the Arabic language in particular is known for its eloquence, conciseness as well as richness. The same word can often have many different meanings—e.g., the Arabic word "*min*" is said to have at least 15 different meanings[9] According to Islamic scholars, the correct understanding of Qur'an can only

[8] http://qurandatabase.org/.

[9] Ibn Hisham's famous *Mughni al-Labib*, an exhaustive compendium of Arabic prepositions and particles, lists fifteen possible meanings of the preposition "*min*" when used in a sentence (Yusuf 2011).

be obtained by knowing both the context (both grammatically and also chronologically in terms of how and when the verse was revealed). For a long time, Qur'an was not translated to other languages because Arabic is assumed to be essential for understanding the Qur'an—even though now many "translations of the meanings" of the Qur'an have been written for pragmatic reasons that dictate the dissemination of the meanings of the Qur'an.

Making a mistake in misunderstanding the nuances of the Qur'an can have grave consequences. In the Islamic tradition, in-depth knowledge of the Arabic language is one of the keys to understanding Islamic principles and laws. In the Islamic tradition, it is commonly assumed that to attempt a translation (*tarjumah*) or an elaboration (*tafsir*) of the Qur'an, one needs to master many sciences including the Arabic grammar (*ilm al-nahw*), morphology (*ilm al-sarf*), rhetoric (*ilm al-balaghah*), the causes of revelation (*asbab al-nuzul*), etymology (*ilm al-ishtiqaq*). There are authentic reports of the Prophet of Islam admonishing those people who speak in matters of the Qur'an from their opinion without due diligence and educational pedigree.[10]

Despite tremendous progress and recent successes, NLP technologies and computer translation still have a long way to go (Hofstadter 2018). This is particularly the case for low-resource languages such as Arabic. Arabic NLP tools particularly for Qur'anic Research need to have humans-in-the-loop and algorithmic explanations and understanding need to be examined and verified by human scholars. Arabic NLP can however still play a strong role. For example, it can facilitate tasks such as search, and information retrieval and help humans in performing these quickly and at scale.

In this section, we highlight some caveats of taking an AI-based NLP approach to Qur'anic research:

1. We should be wary of using the results of these models without involving subject-matter experts, particularly for results that are not aligned with previous scholarly interpretations and results. Any result that goes contrary to the consensus of early scholars should be re-checked and scrapped if it cannot be justified. This is necessary because notwithstanding the great advances in AI-based NLP, and the use of the term "deep" in "deep learning" and "deep neural networks", AI-based NLP solutions do not in general provide any deep human-level understanding and their results remain shallow (Hofstadter 2018). That means, even in the future if we ever had a functioning explainable AI, results must conform to the authoritative scholarly interpretations. In this regard, we are following an early scholar's golden rule, "this knowledge constitutes your religion, so be wary of whom you take your religion from" (Al-Azami 2020, p. 13). This is not to diminish the practical utility of AI-based NLP, and the value of the AI-based NLP models would primarily be in their instrumental value in facilitating quicker research by being more scalable.
2. It is important to develop customized tools and an ecosystem around Qur'anic research. This is because using stock tools from other languages or those tools that have been designed for MSA may not provide reasonable results when applied to Qur'anic applications. This is also needed since mainstream language models trained on unrelated corpora is that they may contain unexpected biases. For instance, Abid et al. (2021) have analyzed the state-of-the-art contextual language model GPT-3 and have shown that it gives persistent bias in tasks such as prompt completion, analogical reasoning,

[10] Mishkat al-Masabih 234 https://sunnah.com/mishkat:234.

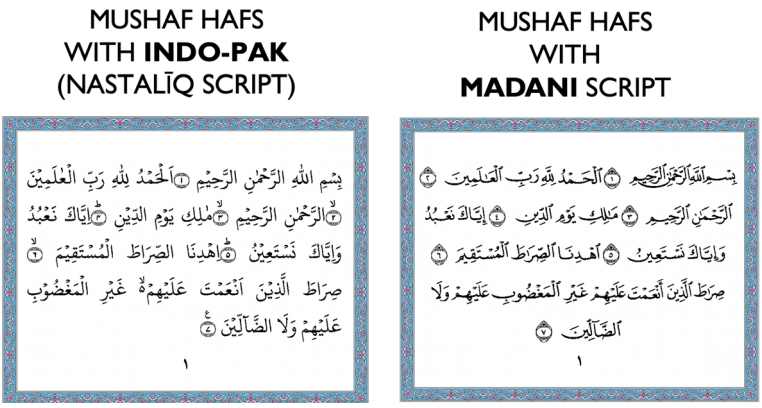


Fig. 15 The first chapter (Surah al-Fatiha) of the Qur'an (Hafs Reading) in South Asian style (Mushaf in Indo-Pak Nasta'liq Script) compared with the Mushaf in Madani script

etc. These problems can be caused by biases in the data and in general ML-based NLP systems replicate and even amplify the biases in the training data and the presence of these errors requires human oversight for applications deploying these models (Jurafsky and Martin 2021). This also underscores the previously-mentioned point that emphasized the need to have the results provided by the models vetted by experts for accuracy keeping in view the sensitive nature of the task of interpreting religious texts.

6 Open issues and future research directions

As compared to English NLP, Qur'anic NLP is still evolving. The important open issues and directions for future research are highlighted next.

1. For Qur'anic research, it is often useful to resort to non-Quranic corpus—in particular, the Hadith corpus and other classical Arabic Islamic books—since Qur'anic Arabic, or Classical Arabic, is different from Modern Standard Arabic and linguistic context of rare (*gharib*) words can be found by consulting the complementary resources. Future Qur'anic NLP research works should consider using these complementary data sources for developing better models and solutions.
2. A researcher working in Qur'anic NLP should be aware of the nuances of Qur'anic orthography and recitation. Prof. Abdel Haleem tracks the history of Qur'anic orthography in (Haleem 1994) and the agreement of Islamic scholars upon *al-rasm al-'Uthmānī* (the way of writing the text of the Qur'ān compiled during the time of the Caliph Uthman b. Affan), also referred to as *rasm al-mushaf*, as the standard written representation of the recited text of the Qur'an. The orthography of the Qur'an differs from the present, and it is hard to pin down to simple rules. For instance, all the 190 occurrences for "the Heavens" and "Heavens" are spelled السموت and سموت respectively, except once it is spelled السموات (Al-Azami 2020, p. 145).

Fig. 16 Qur'anic verse (1:2) provided to the *Farasa* tool (Far) for part of speech tagging resulting in inconsistent results. Poor results are shown for the Qur'an Mushaf with Nasta'liq Script 15. (*Legends*: **S**: Start; **DET**: determiner; **NOUN**: noun; **PREP**: preposition; **NSUFF**: noun suffix; **PUNC**: punctuation; **ABBREV**: abbreviation; and **E**: End.)

3. Very few existing works have considered multimodal NLP models trained on text as well as data from other modalities including audio, and images. It is important to explore what new diverse capabilities can emerge from using multimodal NLP models for Qur'anic research.
4. A lot of partial ontologies have been developed, which cover different concepts from the Qur'an. There is a need to develop a single consistent ontology based on Qur'an and Hadith that covers all the concepts from the Qur'an. Such an ontology can help in knowledge extraction, question answering, and search systems.
5. Techniques based on deep learning have not been much utilized for Qur'anic NLP. Deep learning techniques can be useful in tasks where context remembrance is important and for the Qur'an, context is important while extracting knowledge. Recurrent Neural Networks (RNN) in the form of LSTM and BiLSTM, and other transformer-based architectures can be applied in various Qur'anic NLP tasks. There are a few transformer works emerging (Premasiri et al. 2022; Wasfey et al. 2022), but much more work remains to be done.
6. Intelligent search systems are required for Qur'an, which should be able to answer the queries of people including Muslims and non-Muslims. Most of the current semantic and concept-based search systems are designed from the topic perspective and not from a user perspective.
7. It is also well-known that multiple authenticated readings of the Qur'an have been reported from the Prophet Muhammad (Al-Azami 2020). Such diversity cannot be ignored in any information retrieval process and accounting for this is yet another open issue and future research direction.
8. It is important to note that the Qur'anic scripting method used in the Indo-Pak subcontinent slightly differs from the style used in Arab countries even though they agree upon Al-Rasm al-'Uthmānī (see Fig. 15). The Mushaf in Indo-Pak script relies upon the method of Imam Dani and Imam Ibn Dawud in writing some alphabets and diacritics per the Uthmani script in a way different from their rendering in the Mushaf in Madani script (Ajmal and Lodhi 2018). A major issue with many existing systems is that they are not able to recognize Qur'an text properly when expressed in Indo-Pak script. For instance, a Qur'anic verse (1:2) is provided as input to the online Arabic text processing

tool "Farasa" (Far) for part of speech tagging, both the verses are tagged completely differently (Fig. 16). This highlights the problems that many Arabic NLP tools face when dealing with the different Qur'anic scripts.

7 Conclusions

Qur'anic NLP is an important area of research due to the importance of the Qur'an as the holy book of Muslims, whose global population touches 2 billion people. There has been a lot of recent interest in developing AI-based NLP tools for the Arabic language. However, the research on Qur'anic NLP is less mature compared to Arabic NLP, which is itself compared to a low-resource language for which limited tools and data are available when compared to works focusing on the English language. The challenges related to the Qur'an are quite complex owing to its rich style, its orthography, and most importantly its sacredness. Researchers are trying to overcome the challenges related to all the domains of NLP but more effort is still required. Based on these NLP technologies for Qur'an, many beneficial online applications and systems can be developed that will be helpful for both Muslims and non-Muslims. In this paper, we have provided a comprehensive survey of Qur'anic Arabic focused NLP techniques, tools, and applications. Apart from discussing the various techniques, tools, and applications that researchers have used in past research, we also provide a section on potential pitfalls and caveats and discuss open research issues and highlight promising directions for future work. This is the first comprehensive survey on this important topic and it will be useful as a reference for researchers and practitioners interested in working in this domain.

Funding Open Access funding provided by the Qatar National Library. Not applicable.

Data availability Available on request.

Code availability Not applicable.

Declarations

Conflict of interest No significant conflicts of interests/competing interests.

References

Abid A, Farooqi M, Zou J (2021) Persistent Anti-Muslim Bias in Large Language Models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21). Association for Computing Machinery, New York, NY, USA, 298-306. https://doi.org/10.1145/3461702.3462624.

Abro B, Naqvi AB, Hussain A (2012) Qur'an recognition for the purpose of memorisation using speech recognition technique. In: 15th International Multitopic Conference (INMIC).IEEE 2012. pp 30–34
Afzal H, Mukhtar T (2019) Semantically enhanced concept search of the holy Quran: Qur'anic English wordnet. Arab J Sci Eng 44(4):3953–66
Ahmed AH, Abdo SM (2017) Verification system for Quran recitation recordings. Int J Comput Appl 163(4):6–11
Ahmed R, Atwell E (2016) Developing an ontology of concepts in the Qur'an. Int J Islam Appl Comput Sci Technol 4(4):1–8
Ahsiah I, Noor N, Idris M (2013) Tajweed checking system to support recitation. In: 2013 International conference on advanced computer science and information systems (ICACSIS). IEEE, pp 189–193
Ajmal HM, Lodhi MI (2018) Methods of the Uthmanic Script of the Quran-A Special Study of Pakistani Printed Masahif. Peshawar Islamicus 9(1):1–12
Al-Azami MM (2020) The history of the Qur'anic text from revelation to compilation: a comparative study with the old and new testaments, 2nd edn. Turath Publishing, London
Al-Bakeri AA, Basuhail AA (2017) ASR for Tajweed rules: integrated with self-learning environments. Int J Inf Eng Electr Bus 9(6):1
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
Alhawarat M (2015) Extracting topics from the holy Quran using generative models. Int J Adv Comput Sci Appl 6(12):288–294
Ali BBM, Ahmad M (2013) Al-Quran themes classification using ontology. Icoci Cms Net My 74:383–389
Ali MA, Maged ME (2020) Imam: Word embedding model for Islamic Arabic nlp. In: 2nd novel intelligent and leading emerging sciences conference
Al-Kabi MN, Kanaan G, Al-Shalabi R, Nahar K, Bani-Ismail B (2005) Statistical classifier of the holy Quran verses (fatiha and yaseen chapters). J Appl Sci 5(3):580–583
Al-Khalifa HS, Al-Yahya MM, Bahanshal A, Al-Odah I (2009) Semq: a proposed framework for representing semantic opposition in the holy Quran using semantic web technologies. In 2009 international conference on the current trends in information technology (CTIT)
Alqahtani M, Atwell E (2017) Evaluation criteria for computational Quran search. Int J Islamic Appl Comput Sci Technol 5(1):12–22
Alqahtani M, Atwell E (2015) A review of semantic search methods to retrieve information from the Qur'an corpus. In Corpus Linguistics 2015. Leeds
Alqahtani M, Atwell E (2016) and merging ontology in al-Quran domain. In: 9th Saudi Students conference in the UK. Leeds
Alqahtani M, Atwell E (2016) Arabic Quranic search tool based on ontology. In: International conference on applications of natural language to information systems. Springer, pp 478–485
Alrabiah M, Alhelewh N, Al-Salman A, Atwell E (2014) An empirical study on the holy Quran based on a large classical Arabic corpus. Int J Comput Linguist (IJCL) 5(1):1–13
Alshareef A, El Saddik A (2012) A Quranic quote verification algorithm for verses authentication. In: 2012 international conference on innovations in information technology (IIT). IEEE, pp 339–343
Alshammeri M, Atwell E, Alsalka MA (2020) Qur'anic topic modelling using paragraph vectors. In: Conference: intelligent systems and applications proceedings of the 2020 intelligent systems conference (IntelliSys), vol 2. Springer, pp 218–230
Alsmadi I, Zarour M (2017) Online integrity and authentication checking for Quran electronic versions. Appl Comput Inf 13(1):38–46
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. Arab J Sci Eng 35(2):21
Amazon mechanical turk https://www.mturk.com/
Amrani MYE, Rahman M, Wahiddin MR, Shah A (2016) Towards using CMU sphinx tools for the holy Quran recitation verification. Int J Islamic Appl Comput Sci Technol 4(2):10–5
Atwell E, Habash N, Louw B, Abu Shawar B, McEnery T, Zaghouani W, El-Haj M (2010) Understanding the Quran: a new grand challenge for computer science and artificial intelligence. ACM-BCS Visions of Computer Science
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. Leeds, pp 1–8
Azmi AM, Aljafari EA (2018) Universal web accessibility and the challenge to integrate informal Arabic users: a case study. Univ Access Inf Soc 17(1):131–145

Azmi AM, Almajed RS (2015) A survey of automatic Arabic diacritization techniques. Nat Lang Eng 21(3):477–495
Azmi AM, Alsaiari A (2014) A calligraphic based scheme to justify Arabic text improving readability and comprehension. Comput Hum Behav 39:177–186
Azmi AM, Al-Qabbany AO, Hussain A (2019) Computational and natural language processing based studies of hadith literature: a survey. Artif Intell Rev 52(2):1369–1414
Bentrcia R, Zidat S, Marir F (2018) An analytical study on the holy Quran based on the order of words in Arabic and conjunction. Malays J Comput Sci 31(1):1–16
Brierley C, Sawalha M, Atwell E (2014) Tools for Arabic natural language processing: a case study in qalqalah prosody. In: 9th international conference on language resources and evaluation. European Language Resources Association, pp 283–287
Buckwalter T (2004) Buckwalter Arabic morphological analyzer version 2.0. https://catalog.ldc.upenn.edu/docs/LDC2004L02/readme.txt
Chakroborty S, Roy A, Saha G (2007) Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. Int J Signal Process 2(11):2554–2561
Dimitrakakis C, Bengio S (2011) Phoneme and sentence-level ensembles for speech recognition. Speech Music Process EURASIP J Audio 2011:1–17
Dukes K (2009–2017) The quranic Arabic corpus. https://corpus.quran.com/
Dukes K, Buckwalter T (2010) A dependency treebank of the Quran using traditional Arabic grammar. In: 2010 The 7th international conference on informatics and systems (INFOS), pp 1–7
Dukes K, Buckwalter T (2010) A dependency treebank of the Quran using traditional Arabic grammar. In: 2010 the 7th international conference on informatics and systems (INFOS). IEEE, pp 1–7
Dukes K, Habash N (2010) Morphological annotation of Quranic Arabic. In: LREC
Dukes K, Atwell E, Sharaf A-B (2010) Online visualization of traditional Quranic grammar using dependency graphs. In The Foundations of Arabic Linguistics Conference. Citeseer
Dukes K, Atwell E, Sharaf A-BM (2010) Syntactic annotation guidelines for the Quranic Arabic dependency treebank. In: LREC
Dukes K, Atwell E, Habash N (2013) Supervised collaboration for syntactic annotation of Quranic Arabic. Lang Resour Eval 47(1):33–62
El Amrani MY, Rahman MH, Wahiddin MR, Shah A (2016) Building CMU sphinx language model for the holy Quran using simplified Arabic phonemes. Egypt Inf J 17(3):305–314
ElAffendi MA, Abuhaimed I, AlRajhi K (2021) A simple Galois power-of-two real time embedding scheme for performing Arabic morphology deep learning tasks. Egypt Inf J 22:35–43
ElSayed KN (2015) An Arabic natural language interface system for a database of the holy Quran. Int J Adv Res Artif Intell 4(7):9–14
Ensaf HM, Eyad MS (2022) Qsst: A quranic semantic search tool based on word embedding. J King Saud Univ Comput Inf Sci 34:934–945
Faiza B, Hamid A, Eddine ZD (2021) Semantic query for quranic ontology. J King Saud Univ Comput Inf Sci 33:753–760
Farasa. https://farasa.qcri.org/
Farghaly A, Shaalan K (2009) Arabic natural language processing: challenges and solutions. ACM Trans Asian Lang Inf Process (TALIP) 8(4):14
Fernández-López M, Gómez-Pérez A, Juristo N (1997) Methontology: from ontological art towards ontological engineering. In: Proceedings of the ontological engineering AAAI-97 spring symposium series. American Association for artificial intelligence
Gales M, Young S (2007) The application of hidden Markov models in speech recognition. Now Publisher, Cambridge
Habash NY (2010) Introduction to Arabic natural language processing. Synth Lect Hum Lang Technol 3(1):1–187
Haleem AM (1994) Qur'ānic orthography: the written representation of the recited text of the qur'ān. Islam Q 38(3):171
Hamed SK, Ab Aziz MJ (2016) A question answering system on holy Quran translation based on question expansion technique and neural network classification. JCS 12(3):169–177
Hammo B, Sleit A, El-Haj M (2007) Effectiveness of query expansion in searching the holy Quran. In: The second international conference on Arabic language processing CITALA
Hamoud B, Atwell E (2017) Evaluation corpus for restricted-domain question-answering systems for the holy Quran. Int J Sci Res 6(8):1133–1138
Hanum HM, Bakar ZA, Ismail M (2013) Evaluation of Malay grammar on translation of Al-Quran sentences using earley algorithm. In: 2013 5th international conference on information and communication technology for the Muslim World (ICT4M). IEEE, pp 1–4

Hassan GS, Mohammad SK, Alwan FM (2015) Categorization of 'holy Quran-tafseer' using k-nearest neighbor algorithm. Int J Comput Appl 129(12):1–6
Hofstadter D (2018) The shallowness of Google Translate. https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/
Ibrahim NJ, Razak Z, Yusoff ZM, Idris MYI, Tamil EM, Noor NM, Rahman A, Naemah N (2008) Quranic verse recitation recognition module for support in j-qaf learning: a review. Int J Comput Sci Netw Secur 8(8):207–216
ISO/IEC 9075-1:2008, Information technology - Database languages - SQL - Part 1: Framework (SQL/Framework). https://www.iso.org/standard/45498.html
Iqbal R, Mustapha A, Yusoff ZM (2013) An experience of developing Quran ontology with contextual information support. Multicult Educ Technol J 7(4):333–343
Jamaliah Ibrahim N, Yamani Idna Idris M, Razak Z, Naemah Abdul Rahman N (2013) Automated Tajweed checking rules engine for Quranic learning. Multicult Educ Technol J 7(4):275–287
Jurafsky D, Martin JH (2021) Speech and language processing, 3rd edition (draft). https://web.stanford.edu/~jurafsky/slp3/
Kammani A, Safeena R (2013) Towards a knowledge-based Quran translation: a conceptual model. In: 2013 Taibah University international conference on advances in information technology for the Holy Quran and Its Sciences. IEEE, pp 376–380
Kammani A, Safeena R (2014) A review of Quranic computation for e-learning. Int J Web Sci 2(3):127–139
Khan HU, Saqlain SM, Shoaib M, Sher M (2013) Ontology based semantic search in holy Quran. Int J Future Comput Commun 2(6):570
Malhas R, Elsayed T (2020) Ayatec: building a reusable verse-based test collection for Arabic question answering on the Holy Qur'an. ACM Trans Asian Low-Resour Lang Inf Process (TALLIP) 19(6):1–21
Malhas R, Elsayed T (2022) Arabic machine reading comprehension on the holy Qur'an using CL-AraBERT. Inf Process Manag 59(6):103068
Malhas R, Mansour W, Elsayed T (2022) Qur'an qa 2022: overview of the first shared task on question answering over the holy qur'an. In: Proceedings of the 5th workshop on open-source arabic corpora and processing tools (OSACT5) at the 13th language resources and evaluation conference (LREC 2022)
Mannaa ZM, Azmi AM, Aboalsamh HA (2022) Computer-assisted i'raab of Arabic sentences for teaching grammar to students. J King Saud Univ-Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2022.08.020
Menwa A, Eric A, Mhd A (2021) Detecting semantic-based similarity between verses of the quran with doc2vec. Procedia Comput Sci 189:351–358
Mohamed R, Ragab M, Abdelnasser H, El-Makky NM, Torki M (2015) Al-bayan: a knowledge-based system for Arabic answer selection. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp 226–230
Mohammed A, Sunar MS, Salam MSH (2015) Quranic verses verification using speech recognition techniques. Jurnal Teknologi 73(2):1–8
Mohammed A, Sunar MSB, Salam MSH (2017) Recognition of holy Quran recitation rules using phoneme duration. In: International conference of reliable information and communication technology. Springer, pp 343–352
Muhamad Fahmi F, Moch. Arif B, Arief FH (2020) Implementation of automatic text summarization with textrank method in the development of al-qur'an vocabulary encyclopedia. In: 5th international conference on computer science and computational intelligence
Muhammad AB (2012) Annotation of conceptual co-reference and text mining the qur'an. http://etheses.whiterose.ac.uk/4160/
Muhammad A, ul Qayyum Z, Tanveer S, Martinez-Enriquez A, Syed AZ (2012) E-hafiz: intelligent system to help Muslims in recitation and memorization of Quran. Life Sci J 9(1):534–541
Muhammad MMK, Hassan M, Tengku MTS, Nurhafizah MMY, Sharyar W, Yonis G, Mohd HMH, Syahaneim M, Zahri Y (2021) Semantic graph knowledge representation for Al-Quran verses based on word dependencies. Malays J Comput Sci 31:132–53
Noy NF, Musen MA (2003) The prompt suite: interactive tools for ontology merging and mapping. Int J Hum–Comput Stud 59(6):983–1024
Nur APR, Nurul HAHM (2021) Text categorisation in quran and hadith: overcoming the interrelation challenges using machine learning and term weighting. J King Saud Univ Comput Inf Sci 33:658–667
Overview of the CMUSphinx toolkit https://cmusphinx.github.io/wiki/tutorialoverview/
Palmer M, Gildea D, Kingsbury P (2005) The Proposition Bank: an annotated corpus of semantic roles. Comput Linguist 31(1):71–106

(2004) Protégé https://protegewiki.stanford.edu/wiki/Protege
Palmer M, Babko-Malaya O, Bies A, Diab M, Maamouri M, Mansouri A, Zaghouani W (2008) A pilot Arabic Propbank. In: Proceedings of the sixth international conference on language resources and evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/pdf/880_paper.pdf
Pitchay SA, Ridzuan F (2016) A systematic review analysis for Quran verses retrieval. J Eng Appl Sci 100(3):629–634
Premasiri D, Ranasinghe T, Zaghouani W, Mitkov R (2022) Dtw at qur'an qa 2022: Utilising transfer learning with transformers for question answering in a low-resource domain. arXiv:2205.06025
Putra B, Atmaja B, Prananto D (2012) Developing speech recognition system for Quranic verse recitation learning software. IJID (Int J Inf Dev) 1(2):14–21
Putra SJ, Mantoro T, Gunawan MN (2017) Text mining for Indonesian translation of the Quran: a systematic review. In: 2017 international conference on computing, engineering, and design (ICCED). IEEE, pp 1–5
Qayyum A, Latif S, Qadir J (2018) Quran reciter identification: a deep learning approach. In: 2018 7th international conference on computer and communication engineering (ICCCE). IEEE, pp 492–497
"Qurān" (1999) https://www.britannica.com/topic/Quran
Rubin DL, Knublauch H, Fergerson RW, Dameron O, Musen MA (2005) Protégé-owl: creating ontology-driven reasoning applications with the web ontology language. AMIA
Saad S, Noah SAM, Salim N, Zainal H (2013) Rules and natural language pattern in extracting Quranic knowledge. In: 2013 Taibah University international conference on advances in information technology for the Holy Quran and its sciences. IEEE, pp 381–386
Salloum SA, AlHamad AQ, Al-Emran M, Shaalan K (2018) A survey of Arabic text mining. In: Shaalan K, Hassanien AE, Tolba F (eds) Intelligent natural language processing: trends and applications. Springer, Cham, pp 417–431
Satori H, Harti M, Chenfour N (2007) Introduction to Arabic speech recognition using CMUSphinx system. arXiv:0704.2083
Sharaf A-B, Atwell E (2012) QurSim: A corpus for evaluation of relatedness in short texts. In: Proceedings of the eighth international conference on language resources and evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA), pp 2295–2302. http://www.lrec-conf.org/proceedings/lrec2012/pdf/190_Paper.pdf
Sherif MA, Ngonga A-CN (2015) Semantic Quran. Semantic Web 6(4):339–345
Shoaib M, Yasin MN, Hikmat UK, Saeed MI, Khiyal MSH (2009) Relational wordnet model for semantic search in holy Quran. In: 2009 international conference on emerging technologies. IEEE, pp 29–34
Siddiqui MA, Faraz SM, Sattar SA (2013) Discovering the thematic structure of the Quran using probabilistic topic model. In: 2013 Taibah University international conference on advances in information technology for the Holy Quran and its sciences. IEEE, pp 234–239
SPA (2008) SPARQL Query Language for RDF. https://www.w3.org/TR/rdf-sparql-query/
Ta'a A, Abed Q, Ahmad M (2017) Al-Quran ontology based on knowledge themes. J Fund Appl Sci 9(5S):800–817
Tabbaa HMA, Soudan B (2015) Computer-aided training for Quranic recitation. Procedia Soc Behav Sci 192:778–787
Tabbal H, El Falou W, Monla B (2006) Analysis and implementation of a Quranic verses delimitation system in audio files using speech recognition techniques. In: 2006 2nd international conference on information & communication technologies, vol 2, pp 2979–2984
Wasfey A, Elrefai E, Marwa M, Nawaz H (2022) Stars at qur'an qa 2022: Building automatic extractive question answering systems for the holy qur'an with transformer models and releasing a new dataset. In: Proceedings of the 5th workshop on open-source Arabic corpora and processing tools (OSACT5) at the 13th language resources and evaluation conference (LREC 2022)
Yauri AR, Kadir RA, Azman A, Murad MAA (2012) Quranic-based concepts: verse relations extraction using Manchester owl syntax. In: 2012 international conference on information retrieval & knowledge management. IEEE, pp 317–321
Yekache Y, Mekelleche Y, Kouninef B (2012) Towards Quranic reader controlled by speech. arXiv:1204.1566
Yunus M, Zainuddin R, Abdullah N (2010) "Visualizing Quran documents results by stemming semantic speech query," in *2010 International Conference on User Science and Engineering (i-USEr)*. IEEE, pp. 209–213
Yunus M, Zainuddin R, Abdullah N (2010) Semantic query with stemmer for Quran documents results. In: 2010 IEEE conference on open systems (ICOS 2010). IEEE, pp 40–44

Yusuf HH (2011) The importance of being ambiguous or the sin tax of ignoring syntax. https://tinyurl.com/y7nge9j6
Zaghouani W (2014) Critical survey of the freely available Arabic corpora. In Proceedings of the international conference on language resources and evaluation (LREC'2014), OSACT Workshop. Reykjavik, Iceland, vol 26–31 May. arXiv:1702.07835
Zaghouani W, Dukes K (2014) Can crowdsourcing be used for effective annotation of Arabic? In LREC, pp 224–228
Zaghouani W, Hawwari A, Diab M (2012) A pilot propbank annotation for Quranic Arabic. In: Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature, pp 78–83
Zarrabi-Zadeh H (2007–2021) Tanzil. http://tanzil.net/