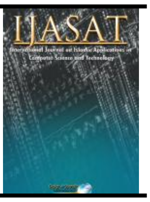

Survey of Semantic Islamic Search Systems

Sanaa Alowaidi [1], Eric Atwel[2] and Mohammad Ammar Alsalka[3]
[1,2,3] University of Leeds, Leeds, UK
[1] King Abdulaziz University, Jeddah, KSA
[1] ml20sara@leeds.ac.uk, [2] e.s.atwell@leeds.ac.uk, [3]m.a.alsalka@leeds.ac.uk

Abstract

A semantic search is a state-of-the-art approach that significantly improves search results since it considers the relationships between the words and the domain concepts rather than only relying on exact word matches. It is clear from the literature that works in semantic search in English are significantly developed compared to Arabic. This paper will highlight the search models focusing on the semantic approaches that leverage the ontology as a knowledge resource. Then, it will narrow the scope to explore the existing Islamic semantic search models to address their limitations and point to possible future directions. As a result of this survey, we notice a severe deficiency in the available resources that interlink varied Islamic sources, such as Quraan and Hadith, and the tools that extract domain-specific concepts. Thus, in the future, a researcher should focus on filling this gap and introducing a public gold standard resource to cover Islamic topics comprehensively for developing Islamic NLP research.

Keywords: Islamic Search Systems, Semantic Search, Knowledge Resource, Ontology, Arabic NLP.

1. Introduction

Over the past decade, several pioneering search systems have emerged, which can be classified into keyword-based and semantic-based. In contrast to the keyword-based search that relies on exact word matches, a semantic search considers the relationships between the words and the domain concepts. The semantic techniques invented focus on enhancing two main tasks of the Question-Answering (QA) system architecture: query analysis and document representation. In the first method, the user query is boosted semantically with several semantic features, such as synonyms, Name Entity (NE), and domain-specific concepts, via natural language processing (NLP) and query expansion techniques (Xiao and Cruz, 2005; Erekhinskaya et al., 2020). Moreover, the second method is building a semantic document representation by incorporating concepts into the representation of the texts (Cao et al., 2008; Ngo & Cao, 2018; Paralic & Kostial, 2003). The proposed methods show an improvement in traditional search systems and aid in understanding user intention and overcoming the keyword mismatching problem. However, semantic models depend on the availability of domain knowledge-based resources to detect concepts and relationships.

The knowledge resources are used to develop computational solutions for several NLP applications, including search systems. Nevertheless, the production quality and quantity of Arabic knowledge resources are still in the early stages compared to English.

This paper aims to review and evaluate the semantic methods that leverage ontology in search systems for Islamic texts to find their limitations and open new directions for future NLP research. To achieve this goal, we will first broaden our scope to review the state-of-the-art

semantic search paradigms proposed and how ontology has been utilized for the English language. Then, we will extensively investigate the available search systems for the Islamic domain and evaluate the available systems against several criteria.
The following sections will first review the English search system in English. Then, the Islamic search models will be presented.

2. Search Systems in the English Language

In this section, we will broaden our focus on literature reviews to address the state-of-the-art search systems approaches in English. First, we will give an overview of the existing search techniques. Then, we will investigate the semantic approaches to incorporate the ontology's semantic features into search systems.

2.1 Search Techniques

Over the years, QA systems have applied two primary approaches: information-retrieval-based and knowledge-based. In addition, these approaches can be distinguished according to the basic representation elements into two types: keyword-based and semantic-based. The keywords-based is entirely dependent on the bag of words as an essential component of the texts (Ruambo & Nicholaus, 2019). Information-retrieval-based QA systems are a well-known example of a keyword-based search. In contrast, the knowledge graph-based QA system is an example of a semantic approach that relies on modeling general or specific domain knowledge by forming the relationships between concepts and entities in a machine-readable format (Erekhinskaya et al., 2020; Zou, 2020).

The language model has recently emerged as a new technique that has proven effective in many NLP tasks (Devlin et al., 2018). Researchers in the question-answering field exploit the nature of the pre-trained language model in encoding many facts during the pretraining phase to answer user queries directly. The following sections will give an overview of these approaches.

2.1.1 Keywords-based systems

The Keyword search systems retrieve related texts by relying on exact word matching. These systems mainly apply Information Retrieval (IR) techniques to retrieve a ranked list related to the given query from a vast amount of data(Jurafsky & Martin, 2019). Initially, the documents and the query are represented in a particular format that summarizes and captures their contents. Then, the relatedness between them is measured to retrieve and rank the relevant documents.
The Vector Space Model (VSM) is the basic paradigm for representing documents in keyword-based systems (Ruambo and Nicholaus, 2019). In this paradigm, the term-documents indexer is created after preprocessing the documents based on unigram word counts and TF-IDF weight. Similarly, the user query is represented as a term-weighted vector, then the similarity between the two vectors is calculated using similarity measures such as cosine similarity.

The simplicity of the representation in the IR models allows for the rapid processing of a vast amount of documents (Ruambo & Nicholaus, 2019). Even though these methods help to retrieve relevant results, the VSM is considered a bag-of-word model where words are treated independently without capturing positional features (Jurafsky & Martin, 2019). In addition, they suffer from a lack of semantic issues, meaning they only retrieve the documents that contain words overlapped precisely with the query's terms based on literal lexical matching rather than word meaning matching. Moreover, it neglects the semantic relations between the terms, which eventually affects the retrieval results. Consequently, it fails to understand the user intention from the query and produces false positive results.

2.1.2 Semantic Knowledge base QA system

The lack of semantics in keywords-based IR models has posed a necessity to invent solutions to improve search results semantically. The revelation of the semantic web motivates researchers to develop an intelligent semantic search approach utilizing ontology as a source of knowledge to capture semantic concepts and implicit meaning in user queries and documents (Asim et al., 2019; Erekhinskaya et al., 2020).

The knowledge graph-based QA model is one of the main paradigms for enhancing search systems semantically. The basic idea of this method is to answer a question written in user language by converting it to a query over structured or semi-structured data (Jurafsky & Martin, 2019). The backbone of this approach is modeling the knowledge as a graph representation consisting of nodes and edges capturing domain entities and relations between them. The most used graph representation format is Resource Description Framework (RDF) triples, in which each triple is formed as (subject, predicate, object) (Erekhinskaya et al., 2020).

In addition, the search technique is based on question translation and answer retrieval tasks. First, it is required to translate the user question into RDF format that matches the ontology schema, and then a SPARQL query is performed to retrieve the answer (Erekhinskaya et al., 2020).

The knowledge graph-based QA systems can be classified according to their application as an open or specific domain (Zou, 2020). The Google knowledge graph is one of the most known open domain models. The knowledge graph result appears as a separate tab in the search window. On the other hand, the second type of knowledge graph depends on capturing the concepts and relations of a specific field domain to compose a semantic research system.

2.1.3 Deep Learning Transformer-Based Language Models

Recently, AI research witnessed the revolution of deep learning transformer-based models that have proven their efficiency in multiple NLP applications, such as name entity (NE), Question answering (QA), and Sentiment Analysis (SA) (Devlin et al., 2018). The Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art transformer model extensively used in QA tasks. It consists of two main phases, namely pretraining and fine-tuning.

The research on the BERT model for the English language is superior to the Arabic language due to the availability of large training datasets. Most recent studies focus on pre-trained BERT on Arabic datasets, producing promising pre-trained language models that can be applied in different NLP tasks, such as AraBERT, CAMeL-BERT, and ArabicBERT.

2.2 How to Leverage Ontology to Enhance the Search Model

The keywords-based search returns a list of ranked documents that match user keywords, but it suffers from a lack of semantically understanding of user intention in the query. The user usually formed their query either as keywords or complete sentences. One of the biggest challenges is to provide precise keywords that match the keywords used by the search systems in indexing the documents (Asim et al., 2019; Ruambo & Nicholaus, 2019).

Based on the literature, semantic systems can be categorized in two directions: knowledge-based QA systems and semantic IR-based systems. The semantic IR-based systems proposed hybrid solutions to adopt the IR model with semantic features. Different strategies are applied to leverage the ontology’s semantic features to enhance the search systems. One direction focuses on enhancing the representation of the documents to be more semantically, either by embedding semantic features into document representation or indexing. Another direction focuses on understanding the user intention and capturing the underlying meaning by applying query expansion techniques. The following sections will highlight some approaches for enhancing traditional search systems.

2.2.1 Document Representation

The traditional bag of words representation of the IR models disregards the semantic relations between terms, eventually affecting the retrieval results (Asim et al., 2019). For example, the synonym words “advantages”, “pros,” and “positive sides” are treated as disjoint words in the counting terms vector, which affects the retrieving accuracy. Thus, several approaches focus on transferring the traditional VSM representation of the texts to encapsulate semantic features. The semantic features utilized for the search models include domain-specific concepts (Paralic & Kostial, 2003), name entity tags, and interrogative word tags (Cao et al., 2008; Ngo and Cao, 2018).In this approach, the semantic features are extracted and embedded into the text representation via augmentation or replacement strategies.

The results show that incorporating the semantic features to retrieve the documents outperforms the basic keyword-based search. In addition, it helps improve the search model’s performance and effectively understand the user’s intention.

2.2.2 Indexing the Corpus

Semantic indexing is a semantic search direction that attempts to utilize ontology in indexing the corpus. The simple indexing technique uses a bag of keywords to generate a descriptor of the document’s content(Ruambo & Nicholaus, 2019).

However, this technique fails to capture the underlying semantic content since it represents only the terms that appeared explicitly in the document. On the other hand, the semantic indexing approach enhances the retrieval process to explore the implicit concepts and capture the underlying relations between the documents(Ahmed-Ouamer & Hammache, 2010; Erekhinskaya et al., 2020).

2.2.3 Query Expansion

The user typically forms a question with keywords different than those used in indexing the documents, reducing the search model’s performance to retrieve related documents. Thus, several works have attempted to enhance the traditional search system over the years by concentrating on understanding the user query. One of the critical NLP techniques that emerged to tackle this issue is query expansion. Asim et al. (2019) indicate that the semantic search based on query expansion focuses on augmenting the user query before the IR process with related terms that share similar meanings. Then, the expanded query is sent to the search engine to retrieve the relevant documents.

One of the leading methods to extract the related terms is to utilize an external knowledge resource. This method extracts related terms by exploring related concepts and semantic relations, such as synonyms and hyponyms, available in WordNet or domain-specific ontology(Erekhinskaya et al., 2020).

The expansion process could be applied automatically using NLP techniques or manually interfering with the user. Xiao and Cruz (2005) propose a model to expand the query using multiple domain-specific ontologies. They implement a semantic framework based on ontology to manage and manipulate files. The query is processed and transformed to RDF format, and then the related files are retrieved by exploring the ontologies concepts and relations.

The query expansion technique has proven its effectiveness in retrieving relevant results for the user. Although it is a simple approach, it depends on the availability of comprehensive knowledge resources representing the target domain to improve the matching process's accuracy.

2.2.4 Semantic Language Model Representation

Although deep learning models have outstanding performance in NLP applications with raw data, they still suffer from poor performance with applications that require external knowledge (Goodwin & Demner-Fushman, 2020). Recently, research focused on enhancing the text representation semantically by embedding knowledge graph concepts in several approaches.
(Zhang et al., 2019) propose an Enhanced Language Representation with Informative Entities (ERNIE) approach to embed the external knowledge data into the language representation model. The pretraining phase is based on an annotated dataset consisting of English Wikipedia texts aligned with Wikidata.

The experimental results show that the proposed approach outperforms the traditional BERT model in two knowledge-driven NLP tasks. In addition, they indicate the importance of creating annotated corpora for real-world tasks to build more extensive pretraining data that lead to better language understanding models.

Goodwin and Demner-Fushman (2020) propose an Ontology-based Semantic Composition Regularization (OSCR) approach to embed the ontological features into the language model. The OSCR approach attempts to augment external knowledge during the BERT pretraining phase to enhance the fine-tuning phase for question-answering datasets. OSCR utilizes ConceptNet 5 as an ontology to embed world and domain knowledge. The OSCR approach is evaluated by pretraining BERT with Wikipedia articles. The resulting pre-trained model is then used to fine-tune different question-answering datasets. The results show that applying the OSCR method to a pre-trained BERT model enhances the accuracy, reaching 33.3%.

Embedding ontological knowledge into language models is still emerging and needs further investigation. However, the proposed approaches depend on knowledge graphs and annotated corpus availability. There are valuable resources for English Languages, such as ConceptNet, WikiData, and DBpedia. In contrast, the Arabic language, and precisely the Islamic field, faces a severe scarcity of resources, and efforts must be concerted to build annotated Islamic sources and KBs as a basis for many NLP applications.

3. Islamic Search Systems

The search systems are fundamental applications for people to retrieve answers and learn Islamic knowledge. The current religion search systems can fall into two primary techniques: keyword-based and semantic search (M. Alqahtani & Atwell, 2015; E. H. Mohamed & Shokry, 2020). Keyword search systems rely on retrieving texts that contain the exact word matching. Thus, these systems suffer from a deficiency of semantics and do not retrieve texts that might contain synonymous words or have the same concepts. In contrast, the semantic search model retrieves the results by considering the relationships between the words and the domain concepts.

The semantic search is considered a challenging task and depends on the availability of quality semantic knowledge resources (E. H. Mohamed & Shokry, 2020). The semantic approaches rely on an external knowledge resource to extract semantic features from a text, such as ontological and WordNet (WN). The basic idea of the semantic search approach is to employ domain-specific ontology or WN to generate all the synonyms and concepts that share the same meaning with the query terms to retrieve more relevant results.

3.1 Semantic Ontology-Based Search Systems

Research in Arabic language and Islamic research has devised several ways to benefit from ontology in building semantic research systems. The main directions of these approaches can be classified as a simple web-based hierarchical tree exploring approach (Abbas, 2009; Kadhim et al., 2015; S. Mohamed et al., 2016), knowledge graph-based search systems (Soediono, 2016; Ullah Khan et al., 2013; Zouaoui & Rezeg, 2021), semantic text representation embedding the concepts with the IR technique approach (Abdelnasser et al., 2015; M. M. A. Alqahtani, 2019; E. H. Mohamed & Shokry, 2020) and ontology-based query expansion. The following section will discuss some of the approaches in detail.

3.1.1 Web-based Hierarchical Tree Exploring

One of the pioneering works that improved the Quranic search systems was proposed by (Abbas, 2009). She developed the Qurany ontology to index the Quranic topics into a hierarchical tree structure. The leaf of the ontology represents a link to the Arabic verses and their English translation, which contains the concept. In addition, Abbas built a web-based search tool based on two main search methods: keyword-based and concept-based. The keyword search allows users to retrieve verses by querying in English or Arabic or mixing words, whereas the concepts-based search is based on the Qurrany ontology, which enables users to explore the main concepts and sub-concepts of the Quran through a static tree structure available in HTML format. The experimental results indicate the importance of incorporating the Quranic topics as concepts to enhance the search accuracy and understanding of Quranic knowledge.

In the same vein, Kadhim et al. (2015) and Mohamed et al. (2016) contributed to building a simple semantic search model for Hadith based on ontology. Mohamed et al. (2016) developed a semantic web search tool for Hadith. The ontology consists of a hierarchy of Hadith concepts and their contents extracted manually from Hadith's topics stored in HTML format. The proposed tool only allowed users to search for a concept by exploring it through tree links. In addition, Kadhim et al. (2015) focused on a specific topic of the Hadith field; they built a simple ontology for "Alsalah" the praying topic from Hadith texts. In addition, a web search interface was built to allow the user to search for Hadith by browsing a hierarchal tree of concepts.
Moreover, Kadhim et al. (2015) improve the semantics of the search process by using Alsalah ontology to annotate a few Hadith texts. Consequently, when the user searches for the term "dawn", the search tool retrieves all Hadith containing the concepts "heat" and "sun", even if they are not explicitly mentioned in the query. The performance of the semantic system outperforms a simple keyword system and achieves about 79% precision and 56% recall. The results indicate the importance of using semantic concepts in the search module; however, the search system and Alsalah ontology are unavailable.

Although this web tree exploring approach contributes to Islamic domain knowledge and allows for exploring concepts and their relations, it depends on browsing the tree content without formulating the query in the user's words.

3.1.2 Knowledge Graph-Based Search Systems

Semantic web technologies represent knowledge in a formal format that allows sharing, reusing, and understanding of the knowledge between users and computers. Recently, one significant direction of the Quranic semantic search is by adopting SW techniques such as Resource Description Framework (RDF) for the knowledge representation and Simple Protocol and RDF Query Language (SPARQL) for retrieving data from ontology (Soediono, 2016; Ullah Khan et al., 2013; Zouaoui & Rezeg, 2021). In this approach, the user query and the documents are annotated with concepts from ontology, and then ontology query language is used to retrieve the related results.

Ullah Khan et al. (2013) developed a Quranic semantic search tool based on semantic web search methods. The search tool retrieves all the verses that mention living creatures utilizing the ontology concepts to answer user queries. They built a domain-specific ontology for animals mentioned in the Quran using Protégé. Then, they retrieved the verses that answered the user query using a SPARQL query language. This method improved the search accuracy by semantically retrieving all verses that implicitly mentioned animal concepts. The authors indicate that the Arabic WN is limited in covering synonyms of Quranic words. Additionally, they recommend enhancing the existing ontology with Quranic semantic relations to cover more Quranic terms in all aspects.

Soediono (2016) built a semantic search system based on ontology and SPARQL queries. Using their previously created medical ontology, they developed an annotated corpus for Hadith texts with medical concepts in RDF format. In this system, the search words are extracted from the user question before being transformed into SPARQL queries to retrieve search results. Then, the SPARQL search engine utilized the resulting annotations to retrieve the most related search results. The system focuses on animal concepts and does not cover other knowledge domains. In addition, the query is initiated using SPARQL syntax only, and the system did not support forming a query in natural language text.

Zouaoui and Rezeg (2021) argue that discovering the links between the words in the verses using “Erab” grammar features is vital to understanding queries and semantically retrieving the most related verses. They proposed Quran Erab ontology to represent the morphological and syntax features of Quranic words using OWL. In addition, the Quran corpus is annotated using Erab ontology to capture the semantic links between the words in the same verses. The search model is built using the Apache Jena platform and SPARQL queries. The system allows the user to input a query as free text consisting of one or more words. Then, the query is preprocessed and analyzed to understand the grammatical concepts before being converted to the SPARQL query syntax. Finally, the system applies the SPARQL queries on Erab ontology to retrieve the most semantically related verses. The performance of the proposed semantic search system is evaluated and compared with the keyword search systems. The results show that the proposed method helps to improve the precision of the semantic Quranic search systems. Although the syntax features tagging are essential to analyze the query, it depends on the capability of the morphological tool, which may misunderstand user words. Another shortcoming of this study is that the ontology and the system are not publicly published.

3.1.3 Integrating Concepts with IR Techniques

Recently, research has focused on developing approaches for building a semantic question-answering system by integrating the concepts with several IR techniques (Abdelnasser et al., 2015; M. M. A. Alqahtani, 2019). In this approach, semantic concepts are utilized in the text representation model. Then, IR techniques such as similarity measures or pattern matching are applied to retrieve the related documents.

Abdelnasser et al. (2015) built a semantic QA system called Albayan that semantically retrieves the candidate answers to the user query. The answer comprises the verses and descriptions from Ibn Kathir's Tafsir book. They manually integrated 1200 concepts from two popular ontologies, the QAC and Qurany, which are utilized to annotate the questions with their extracted concepts. Albayn's system goes through three primary phases: question analysis, information retrieval, and answer extraction. In the question analysis phase, each word in the user input question was annotated with different features, including part of speech (POS), word stem, and Named Entity (NE), using the MADA[1] morphological tool and the LingPipe[2] (NER) tool. In addition, the question type was classified automatically using a support vector machine (SVM) ML algorithm. After that, the preprocessed question was passed to the information retrieval phase to retrieve the top semantically related verses by applying the explicit semantic analysis method proposed by (Gabrilovich et al., 2007). This approach is based on extending the keyword text representation with concept features extracted from the external Quranic ontology. In this stage, the Quranic verses and the user query were represented with weighted vectors of concept features. Finally, the cosine similarity was computed between the query vector and the verses vector to extract the most semantically related verses. The results show an improvement in retrieving accuracy, reaching 65%.

Alqahtani (2019) built a semantic model for Quranic search based on ontology. He developed a new ontology to expand the coverage of the concepts by integrating the available Qurany, Arabic Quran Corpus (AQC), and QuraAna resources (M. M. Alqahtani & Atwell, 2018). In addition, he built an Arabic question and answer corpus (AQAC) containing 2224 questions about the Quran collected from an Islamic book and website. The corpus is annotated with multiple morphological, structural, and concept features. The proposed QA system goes through several phases before answering the user question. First, the query was preprocessed based on POS tagging and Arabic stemming utilizing the Stanford CoreNLP[3] and ISRI[4] Arabic Stemmer from (NLTK) toolkit. Then, the resulted words are extended with their synonym generated by apply the word embedded technique using the word2vec algorithm. Next, the cosine similarity was applied to extract concepts from the ontology that match the query and their related verses were relatives. The system comprises a classification layer to classify the question class using the FastText[5] tool, the classifier helps to predict the two most related answer type with about 65.5 % precision. Finally, the results were filtered and ranked based on the number of matched words and concepts between the question and the retrieved answers.

3.1.4 Query Expansion in Arabic QA systems

The query expansion task improves the retrieving performance by extending the user query with synonyms and implicit concepts to capture the hidden words that might not appear in the question. The ontology has been extensively utilized to enrich the user query with their related concepts using general domain ontology such as WN or domain-specific ontology. This section will review studies exploring query expansion methods in Arabic QA systems to improve retrieving and ranking results.

[1] http://www.cs.columbia.edu/~rambow/software-downloads/MADA_Distribution.html

[2] http://www.alias-i.com/lingpipe/demos/tutorial/ne/read-me.html

[3] https://stanfordnlp.github.io/CoreNLP/

[4] https://www.nltk.org/_modules/nltk/stem/isri.html

[5] https://fasttext.cc/

Abouenour et al. (2010) indicate that the performance of the QA system is enhanced by expanding the queries with their related terms using AWN. The keyword-based search is compared against the query expansion model using an Arabic translation version of questions from TREC and CLEF datasets. Similarly, Al-Chalabi et al. (2015) state that semantic enrichment of the user query by relevant keywords is vital to increase the accuracy of the QA systems. They expanded the user query with related terms from AWN ontology. Then, an equivalent semantic question is generated by combining the resulting terms using AND and OR operators. The Google search engine was utilized to retrieve results for 50 questions selected from TREC and CLEF datasets. The experiment results show an enhancement in ranking the retrieved documents by applying QE.

Moreover, Al-Khawaldeh (2019) implemented a QA system called AWAQ for Arabic. The system expands the query and the passage with a synonym from AWN to enhance retrieval accuracy. The answers are retrieved using several search engines. Then, the similarity entailment method based on AWN is applied between the question and the passages to rank the retrieved results. The test dataset contains 250 questions about religion, science, history, computers, and politics. The experiment results demonstrate the effectiveness of query expansion in the QA search system.

3.2 Deep learning QA Systems

The deep learning transformer model is a powerful tool that has proven its effectiveness in multiple NLP tasks. The research on the Arabic QA task is still in the early stages compared to the English language due to the deficiency of NLP resources and datasets for the QA task. Recently, several studies have tended to enrich the Arabic QA resources and apply transformer models such as BERT for Arabic question-answering tasks.

Mozannar et al. (2019) created an Arabic Reading Comprehension Dataset (ARCD) for a QA task comprising 1395 MSA questions from Wikipedia pages and about 2966 translated questions from the SQuAD 1.1 QA English corpus. The proposed corpus is used to train the BERT model for the QA task. The results achieve about 0.61 in the F1 score. In addition, Antoun et al. (2020) developed one of the most used transformer models called AraBERT, an Arabic version of the BERT model trained on 70 million sentences available in Arabic corpora and news websites. The AraBERT model is fine-tuned for the QA task using Arabic-SQuAD and ARCD datasets. The AraBERT outperforms the multilingual BERT (mBERT) by 1.4% in the F1 score.

Additionally, the performance of different transformer models, including AraBERTv2-base and AraBERTv0.2-large, is compared for the QA task (Alsubhi et al., 2021). The models are tested on Arabic QA datasets, such as TyDiQA-GoldP, SQuAD, ARCD, and AQAD, which are mainly based on Wikipedia documents. The AraBERTv0.2-large fine-tuned on the TyDiQA-GoldP dataset achieved the best results, reaching 86.49% in F1. Moreover, the results show a variant performance of the transformer model depending on the dataset. The authors observe that the dataset’s quality strongly affects the model’s accuracy. The performance is increased if the dataset is correctly labeled and cleaned.

The current Arabic QA studies focus on Wikipedia, news, and web forums in different domains. The outstanding performance of the transformer models attracts the researcher to evolve Islamic research. Malhas and Elsayed (2020) created a Quranic QA corpus, which was extended later to produce the ORCD corpus. The ORCD is a QA corpus where answers extracted from Quranic verses consist of 1337 questions and answers. Several studies were conducted to adopt transformer models on top of the ORCD corpus to boost Quranic research under the Quranic 2022 shared task (Malhas et al., 2022). To evaluate the result, they proposed the pRR evaluation metric that considers the results that partially matched the gold answers.

According to Malhas et al. (2022), the proposed approaches adopt several transformers-based language models already pre-trained with MSA only, CA only, or combined MSA, CA, and dialect datasets. The top results are achieved using language models pre-trained with MSA-only. In addition, several studies focus on fine-tuning to enhance the performance of the pre-trained language models. They propose to use the ORCD corpus in combination with other QA datasets to fine-tune the transformer model in a pipelined approach. The experimental results demonstrate that fine-tuning the transformer models with a large dataset enhances their performance. The best result reaching 0.586 in pRR is achieved using the AraELECTRA model that was pre-trained on MSA only and fin-tuned with a multilingual TyDiQA dataset before applying the ORCD dataset. However, using only the ORCD datasets for transformer models' fine-tuning phase has achieved the second-highest pRR score, reaching 0.567 after applying several postprocessing methods to improve answer predicting.

There is no doubt that the transformer model has proven its efficiency; however, it has encountered some difficulties in learning the context of some words related to domain-specific fields. Alsaleh et al. (2021) noticed that the AraBERT model failed to predict the relatedness between two verses when they contained different words but shared the same concepts in Islamic teaching. Moreover, the transformer model failed to answer some questions when there were no match words in the question, even though there exist phrases that share a similar meaning (A. Alsaleh et al., 2022). These results emphasize the importance of enhancing the models with semantic features to capture the underlying concepts, particularly for Islamic knowledge. Furthermore, due to the deficiency of the specialist QA dataset, the research progress is still slow and needs more effort to fill the gap.

3.3 Evaluation of the Current Studies

The semantic search models allow returning an answer by considering the relationships between the words and the domain concepts. Based on our review, one can conclude that the semantic enhancement techniques target the two main tasks of the QA system architecture: query analysis and document representation.

There are few studies introducing various semantic models for Islamic texts. Abbas (2009), Mohamed et al. (2016), and Kadhim et al. (2015) created a knowledge resource that provides a simple ontology exploring method as a tree hierarchical structure. However, it lacks semantic analysis to understand the user query. Moreover, Research by Ullah Khan et al. (2013), Soediono (2016), and Zouaoui and Rezeg (2021) implement knowledge graph QA search models based on SPARQL query to retrieve answers to domain-specific questions. In addition, Abdelnasser et al. (2015) and Alqahtani (2019) utilized ontology as an external knowledge resource to enhance the IR-based search system by embedding the concepts via semantic document representation or query expansion techniques.

This paper evaluates the current semantic Islamic search approaches according to several criteria, as demonstrated in Table 1. The results of the evaluation show severe deficiencies in the existing models. One of the notable drawbacks is that most of the models only cover one source, either the Qur'an or Hadith. Moreover, as mentioned in (Alowaidi et al., 2023), the avaliable Quranic resource covers linking abstract topics with their related verses and neglects the detailed information available in other Islamic resources about the domain-specific relations between concepts and their entities. As a result, retrieving answers to questions mentioned implicitly in the Quran but explicitly found in Hadith is considered a challenging task that will affect the accuracy of the search results. These limitations could be solved by integrating and expanding the existing ontology with more Islamic resources.

The second weakness is a lack of semantic analysis of the query to understand user intention. Most previous works do not provide advanced NLP techniques to analyze the query. They apply basic text preprocessing such as word stem and lemma, allowing users to search by simple keywords and phrases. Abdelnasser et al. (2015) and Alqahtani (2019) propose models that accept prompt a question in a natural language besides applying basic and advanced NLP analysis such as different classification techniques; however, there is a need to embed more semantic features such as NE recognition, synonym features and concepts linking. In the future,

Table 1 Comparison of different Islamic search systems investigated in Section 3

Paper	**Domain**	**Resources**	**Answer type**	**Avali-able**	**Query Type**	**Query analysis**	**Search techniques**
(Abbas, 2009)	Quran	Qurany ontology links the topics with related verses	Verses	Y	One word, two words	-	Simple exploring the ontology as a tree hierarchical structure indexing topics.
Ullah Khan et al. (2013)	Quran, animals	A domain-specific resource annotates verses with related animal entities.	Verses	N	Question	Word stem, word lemma, Convert natural language query to SPARQL	Retrieve the results using the SPARQL query to the ontology.
Abdelnasser et al. (2015)	Quran	A QA corpus consisting of verses, descriptions, and related concepts. It integrates QAC and Qurany.	Verses – descript-ions	N	Question	POS, word stem, word lemma, rule-baes method to classify Question Type	IR-based techniques applying semantic VSM document representation-cosine semantic similarity measures
Kadhim et al. (2015)	Hadith, prayer	A domain-specific resource annotates Hadith texts with prayer concepts	Hadith	N	One word, two words	-	Simple exploring the ontology as a tree hierarchical structure
Mohamed et al. (2016)	Hadith	Hadith ontology links the topics with related Hadith	Hadith	N	One word, two words	-	Simple exploring the ontology as a tree hierarchical structure indexing topics.
Soediono (2016)	Hadith, medical concepts	A domain-specific resource annotates Hadith texts with medical concepts.	Hadith	N	Sentence	Word stem- Ngram-synonym features- Convert natural language query to SPARQL.	Apply SPARQL query to retrieve answers ontology.
Alqahtani (2019)	Quran	AQAC ontology integrates several ontologies and links concepts to their related verses	Verses	N	Question	POS tagging, word stem - Query expansion with synonym generated using the word2vec algorithm	IR-based techniques applying semantic VSM document representation - cosine semantic similarity measures
Zouaoui and Rezeg (2021)	Quran, Erab	A domain-specific Erab ontology represents the morphological and syntax features of Quranic words.	Verses	N	One word, two words	Morphological analysis - synonym features- Convert natural language query to SPARQL.	Apply SPARQL query to reasoning the ontology.

there is a severe need to develop a query analyzer tool that accepts a question in natural language and produces a semantic query to increase understanding of user intention.

4. Conclusion

This paper highlights the state-of-the-art semantic search techniques that leverage ontology as a knowledge resource. It is clear from the previous studies that ontology plays a significant role in most of the proposed solutions. In addition, it has been utilized to enhance search systems in two ways: to generate a semantic document representation capturing the underlying semantic concepts and relations or as an external knowledge resource for semantic query analysis.

On the other hand, this paper introduced an overview of current Islamic semantic search systems and highlighted the main drawbacks and how they could be overcome. It is clear from previous studies that there is a lack of available resources that interlink different Islamic resources and cover Islamic topics comprehensively. In addition, there is a lack of semantic analyzer tools for extracting Islamic concepts. Thus, a public gold standard resource is urgently demanded to cover Islamic topics comprehensively. In the future, Islamic research should focus on developing a resource interlinking Islamic knowledge from the Quran and Hadith. Several existing knowledge resources can be integrated into a unified format by converting all to RDF and then merging equivalent or similar concepts from different resources.

References

Abbas, N. (2009). *Quran ' Search for a Concept ' Tool and Website Quran ' Search for a Concept ' Tool and Website*. *July 2009*. https://www.researchgate.net/profile/Noorhan-Abbas-2/publication/318226723_Quran_'Search_for_a_Concept'_Tool_and_Website/links/596200280f7e9b81946b192e/Quran-Search-for-a-Concept-Tool-and-Website.pdf

Abdelnasser, H., Ragab, M., Mohamed, R., Mohamed, A., Farouk, B., El-Makky, N., & Torki, M. (2015). *Al-Bayan: An Arabic Question Answering System for the Holy Quran*. 57–64. https://doi.org/10.3115/v1/w14-3607

Abouenour, L., Bouzouba, K., & Rosso, P. (2010). An evaluated semantic query expansion and structure-based approach for enhancing Arabic question/answering. *International Journal on Information and Communication Technologies*, *3*(3), 37–51.

Ahmed-Ouamer, R., & Hammache, A. (2010). Ontology-based information retrieval for e-Learning of computer science. *2010 International Conference on Machine and Web Intelligence*, 250–257. https://doi.org/10.1109/ICMWI.2010.5648113

Al-Chalabi, H., Ray, S., & Shaalan, K. (2015). Semantic Based Query Expansion for Arabic Question Answering Systems. *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, 127–132. https://doi.org/10.1109/ACLing.2015.25

Al-Khawaldeh, F. T. (2019). Answer extraction for why Arabic questions answering systems: EWAQ. *ArXiv Preprint ArXiv:1907.04149*.

Alowaidi, S., Atwell, E., & Alsalka, M. (2023). Islamic Ontology Coverage Evaluation. *International Journal on Islamic Applications in Computer Science And Technology*, *11*(2). http://www.sign-ific-ance.co.uk/index.php/IJASAT/article/view/2593

Alqahtani, M., & Atwell, E. (2015). A Review of Semantic Search Methods to Retrieve Information from the Qur ' an Corpus. *8th International Corpus Linguistics Conference*, *July*, 365–368.

Alqahtani, M. M. A. (2019). *Quranic Arabic Semantic Search Model Based on Ontology of Concepts*. *January*. http://etheses.whiterose.ac.uk/24184/

Alqahtani, M. M., & Atwell, E. (2018). Developing Bilingual Arabic-English Ontologies of Al-Quran. *2nd IEEE International Workshop on Arabic and Derived Script Analysis and*

Recognition, ASAR 2018, 96–101. https://doi.org/10.1109/ASAR.2018.8480237

Alsaleh, A., Althabiti, S., Alshammari, I., Alnefaie, S., Alowaidi, S., Alsaqer, A., Atwell, E., Altahhan, A., & Alsalka, M. A. (2022). LK2022 at Qur ' an QA 2022 : Simple Transformers Model for Finding Answers to Questions from Qur ' an. *The 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, June*, 120–125. https://osact-lrec.github.io

Alsaleh, A. N., Atwell, E., & Altahhan, A. (2021). Quranic Verses Semantic Relatedness Using AraBERT. *Proceedings of the Sixth Arabic Natural Language Processing Workshop, 3*, 185–190. https://aclanthology.org/2021.wanlp-1.19

Alsubhi, K., Jamal, A., & Alhothali, A. (2021). Pre-trained transformer-based approach for Arabic question answering: A comparative study. *ArXiv Preprint ArXiv:2111.05671*.

Antoun, W., Baly, F., & Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *ArXiv Preprint ArXiv:2003.00104*.

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, N., & Mahmood, W. (2019). The use of ontology in retrieval: a study on textual, multilingual, and multimedia retrieval. *IEEE Access*, *7*, 21662–21686.

Cao, T. H., Le, K. C., & Ngo, V. M. (2008). Exploring combinations of ontological features and keywords for text retrieval. *Pacific Rim International Conference on Artificial Intelligence*, 603–613.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Erekhinskaya, T., Tatu, M., Balakrishna, M., Patel, S., Strebkov, D., & Moldovan, D. (2020). Ten Ways of Leveraging Ontologies for Rapid Natural Language Processing Customization for Multiple Use Cases in Disjoint Domains. *Open Journal of Semantic Web (OJSW)*, *7*(1), 33–51. http://nbn-resolving.de/urn:nbn:de:101:1-2020112218332779310329

Gabrilovich, E., Markovitch, S., & others. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *IJcAI*, *7*, 1606–1611.

Goodwin, T. R., & Demner-Fushman, D. (2020). Enhancing question answering by injecting ontological knowledge through regularization. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, *2020*, 56.

Jurafsky, D., & Martin, J. H. (2019). *Speech and language processing (3rd (draft) ed.)*. Stanford Univ.

Kadhim, R. J., Norwawi, N. M., Abdulaaziz, A. M., & Al, A. (2015). Extraction of Hadith Based on Semantic Annotation. *IJCSN International Journal of Computer Science and Network*, *4*(2), 2277–5420. www.IJCSN.org

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.*, *19*(6). https://doi.org/10.1145/3400396

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. *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5) at the 13th Language Resources and Evaluation Conference (LREC 2022)*.

Mohamed, E. H., & Shokry, E. M. (2020). QSST: A Quranic Semantic Search Tool based on word embedding. *Journal of King Saud University - Computer and Information Sciences*, *xxxx*. https://doi.org/10.1016/j.jksuci.2020.01.004

Mohamed, S., Hassan, O., & Atwell, E. (2016). Concept Search Tool for Multilingual Hadith Corpus. *International Journal of Science and Research (IJSR)*, *5*(4), 1326–1328. https://doi.org/10.21275/v5i4.nov162788

Mozannar, H., Hajal, K. El, Maamary, E., & Hajj, H. (2019). Neural Arabic question answering. *ArXiv Preprint ArXiv:1906.05394*.

Ngo, V. M., & Cao, T. H. (2018). Ontology-based query expansion with latently related named entities for semantic text search. *ArXiv Preprint ArXiv:1807.05579*.

Paralic, J., & Kostial, I. (2003). Ontology-based information retrieval. *Proceedings of the 14th International Conference on Information and Intelligent Systems (IIS 2003), Varazdin, Croatia*, 23–28.

Ruambo, F. A., & Nicholaus, M. R. (2019). Towards enhancing information retrieval systems: A brief survey of strategies and challenges. *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 1–8.

Soediono, B. (2016). An Ontology Based Approach to Enhance Information Retrieval from Maktabah Shamilah. *Journal of Chemical Information and Modeling*, *53*(1), 160.

Ullah Khan, H., Muhammad Saqlain, S., Shoaib, M., & Sher, M. (2013). Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*, *2*(6), 570–575. https://doi.org/10.7763/ijfcc.2013.v2.229

Xiao, H., & Cruz, I. F. (2005). A Multi-Ontology Approach for Personal Information Management. *Semantic Desktop Workshop*, *8*, 10–12.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). {ERNIE}: Enhanced Language Representation with Informative Entities. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–1451. https://doi.org/10.18653/v1/P19-1139

Zou, X. (2020). A survey on application of knowledge graph. *Journal of Physics: Conference Series*, *1487*(1), 12016.

Zouaoui, S., & Rezeg, K. (2021). A Novel Quranic Search Engine Using an Ontology-Based Semantic Indexing. *Arabian Journal for Science and Engineering*, *46*(4), 3653–3674. https://doi.org/10.1007/s13369-020-05082-5