# Al-Bayan: A Knowledge-based System for Arabic Answer Selection

**Reham Mohamed**
reham.mohmd@alexu.edu.eg
**Maha Ragab**
maha.ragab@alexu.edu.eg
**Heba Abdelnasser**
heba.abdelnasser@alexu.edu.eg
**Nagwa M. El-Makky**
nagwamakky@alexu.edu.eg
**Marwan Torki**
mtorki@alexu.edu.eg
Computer and Systems Engineering Department
Alexandria University, Egypt

## Abstract

This paper describes Al-Bayan team participation in SemEval-2015 Task 3, Subtask A. Task 3 targets semantic solutions for answer selection in community question answering systems. We propose a knowledge-based solution for answer selection of Arabic questions, specialized for Islamic sciences. We build a Semantic Interpreter to evaluate the semantic similarity between Arabic question and answers using our Quranic ontology of concepts. Using supervised learning, we classify the candidate answers according to their relevance to the users questions. Results show that our system achieves $74.53\%$ accuracy which is comparable to the other participating systems.

## 1 Introduction

With the increase of the popularity of community question answering (CQA) systems, answer selection became more challenging. CQA systems are often open for public to answer any questions with no restriction or review from field experts. This highlights the importance of developing systems that automatically detects the most relevant answers from the irrelevant ones. These systems might be open-domain or closed-domain, causing a tradeoff between accuracy and generality.

SemEval-2015 task 3 targets semantically oriented solutions for answer selection in community question answering data. We focus on Subtask A for the Arabic language which provides questions and several community answers from the Fatwa website[1]. The goal is to classify each answer as: Direct, Related or Irrelevant.

In this paper, we propose a knowledge-based answer selection system for Arabic. We use our Quranic ontology, enriched with Quran verses and Tafseer books, to convert each question and its candidate answers into weighted vectors of ontology concepts. We use these vectors to compute a semantic similarity score between the question and each candidate answer. We also compute a keyword matching score and feed the two scores into a decision tree classifier which predicts how much the answer is related to the question.

The rest of the paper is organized as follows: Section 2 shows some of the related work to the system. Section 3 shows the details of the system architecture. In Section 4, we show the results of the task evaluation. Finally, we conclude the paper in Section 5.

## 2 Related Work

Our work is related to prior work in both Quranic research and Question Answer Selection systems.

**(a) Quranic Research:** Several studies have been made to understand the Quranic text and extract knowledge from it using computational linguistics. Saad et al. (2009) proposed a simple methodology for automatic extraction of concepts based on the Quran in order to build an ontology. In (Saad et al., 2010), they developed a framework for automated generation of Islamic knowledge concrete concepts that exist in the holy Quran. Qurany (Abbas, 2009) builds

[1] Fatwa is a question about the Islamic religion.



*Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 226–230,
Denver, Colorado, June 4-5, 2015. 

a Quran corpus augmented with a conceptual ontology, taken from a recognized expert source 'Mushaf Al Tajweed'. Quranic Arabic Corpus (Atwell et al., 2011) also builds a Quranic ontology of concepts based on the knowledge contained in traditional sources of Quranic analysis, including the sayings of the prophet Muhammad (PBUH), and the *Tafseer* books. Khan et al. (2013) developed a simple ontology for the Quran based on living creatures including animals and birds that are mentioned in the Quran in order to provide Quranic semantic search. AlMaayah et al. (2014) proposed to develop a WordNet for the Quran by building semantic connections between words in order to achieve a better understanding of the meanings of the Quranic words using traditional Arabic dictionaries and a Quran ontology.

Other attempts for text-mining the Quran were proposed such as: QurAna (Sharaf and Atwell, 2012) which is a corpus of the Quran annotated with pronominal anaphora and QurSim (Sharaf and Atwell, 2012) which is another corpus for extracting the relations between Quran verses.

**b) Question Answer Selection Systems:** Few attempts have been proposed for Arabic Answer Selection. In CLEF 2012, the Arabic language was introduced for the first time for selecting answers to questions from multiple answer choices of short Arabic texts. Abouenour et al. (2012) proposed a system based on distance density N-gram model and Arabic WordNet expansion. Trigui et al. (2012) proposed another system that used inference rules on the CLEF background collection. However, those systems have low accuracy, 0.21 and 0.19 respectively. In CLEF 2013, Al-QASIM system (Ezzeldin et al., 2013) was proposed which focused on answer selection and validation. This approach divided the task into 3 phases: (i) Document analysis, (ii) locating questions and answers and (iii) answer selection. The overall accuracy of the system is 0.36.

## 3 System Architecture

### 3.1 System Overview

The system architecture is shown in Figure 1. The dataset consists of Arabic questions and their candidate answers. The goal is to classify each candidate answer into: (Direct, Related or Irrelevant).

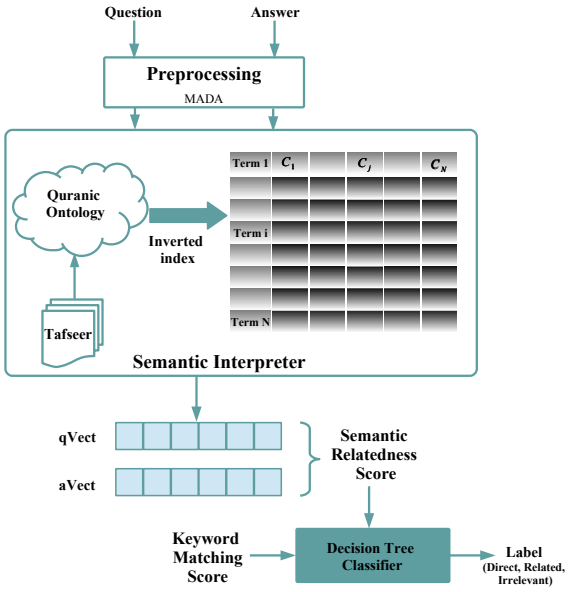


Figure 1: System Architecture.

The question and the answers are preprocessed and fed into the Semantic Interpreter. The Semantic Interpreter uses a Quranic ontology of concepts enriched with Quran interpretation *(Tafseer)* books to build an inverted index. The question is converted into a weighted vector of concepts (qVect) and similarly the candidate answer (aVect). A semantic relatedness score and a keyword matching score are computed and fed into a decision tree classifier which outputs the label of the answer.

### 3.2 Preprocessing

First, we apply morphological analysis on the Arabic text to identify its structure and remove the unwanted words (stopwords). For this purpose, we use MADA (Morphological Analysis and Disambiguation for Arabic) (Habash et al., 2009) which is one of the most accurate Arabic preprocessing toolkits. MADA can derive extensive morphological and contextual information from raw Arabic text, and then use this information for high-accuracy part-of-speech tagging, diacritization, lemmatization, disambiguation, stemming, and glossing in one step.

Each term in the input text is represented by its stem and POS tag using Buckwalter transliteration (Buckwalter, 2002). We identify the stopwords ac-

cording to their POS tags. Pronouns, prepositions, conjunctions and other POS types are all removed.

### 3.3 Building the Ontology

We integrated the Quranic Corpus Ontology (Atwell et al., 2011) and the Qurany Ontology (Abbas, 2009), to form our Quranic conceptual ontology proposed in (Abdelnasser et al., 2014). The **Quranic Corpus Ontology** uses knowledge representation to define the key concepts in the Quran, and shows the relationships between these concepts using predicate logic. The **Qurany Ontology** is a tree of concepts that includes all the abstract concepts covered in the Quran. It is imported from 'Mushaf Al Tajweed' list of topics. This integration was difficult since we had to resolve the overlapping between the two ontologies. There were also some mistakes in the Qurany Concept Tree. So, we had to manually revise the 1200 concepts and their verses.

The Holy Quran consists of 6236 verses. Each verse has to be under at least one concept in our Quranic ontology. After the previous integration process, there were 621 verses without concepts, so we added them under their most suitable concepts to complete the ontology using a similarity measure module. This module measures the similarity between classified and unclassified verses to determine the concepts of unclassified verses. Now, our final ontology contains 1217 leaf concepts and all verses of the Quran. Under each concept in our ontology, we save the related verses with their *Tafseer*, that is used to build the inverted index. We use two *Tafseer*[2] books: (Ibn-Kathir, 1370) and (Al-Jaza'iri, 1986), which are two of the most traditional books used by Islamic scholars. It is possible to add other books to enrich our corpus data.

### 3.4 Building the Semantic Interpreter

We use machine-learning techniques to build a Semantic Interpreter using the Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007) approach. The Semantic Interpreter maps the input Arabic text into a weighted vector of Quranic concepts.

For each leaf concept $C_i$, we construct a document $D_i$ such that $D_i$ contains all the verses related to this concept and their Tafseer. We used Lucene Indexer[3] to build an inverted index on the constructed documents where each term $T_j$ is represented as a weighted vector of concepts. Entries of this vector are assigned weights using the TFIDF scheme which quantifies the strength of association between terms and concepts.

Any input query to the system can be represented as a weighted vector of concepts by calculating the mean of concept vectors of the query terms.

### 3.5 Semantic Relatedness Score

In order to evaluate the semantic relatedness between two Arabic texts, we enter each text into the Semantic Interpreter as a query. The Semantic Interpreter represents each text as a weighted vector of concepts. We compute the Cosine similarity between the two weighted vectors which represents the semantic relatedness score. Therefore, if two texts are semantically related, they will have similar weights for the same concepts and consequently a high Cosine similarity score, and vice versa.

### 3.6 Keyword Matching Score

In this mechanism, the answers of a question are weighted based on the matched words between the answers and the question. For answer $k$ and question term $j$, $Score_{k_j}$ is the number of $j$ repetitions in $k$ normalized by the maximum number of repetitions of $j$ in all answers. $Score_k$ is the summation of $Score_{k_j}$, $(j = 1, .., n)$ where $n$ is the number of the question terms. Finally, we normalize all answers by the maximum $Score_k$.

### 3.7 Answer Classification

We compute the semantic relatedness score and the keyword matching score for each combination of question and answer in the training data. The two scores are normalized for each question. Now to classify the answers as (Direct, Related, Irrelevant), we train a decision tree classifier using the two normalized scores with the gold-standard labels supplied with the training data. The normalized scores are also computed for the test data and the classifier predicts the label of each answers. Results are shown in the next section.

[2]Tafseer is the interpretation of the Quran.

[3]http://lucene.apache.org/

| Class | Direct | Related | Irrelevant | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| Direct | 150 | 40 | 25 | 0.721 | 0.698 | 0.709 |
| Related | 43 | 94 | 85 | 0.519 | 0.423 | 0.467 |
| Irrelevant | 15 | 47 | 502 | 0.820 | 0.890 | 0.854 |
| Macro | - | - | - | 0.687 | 0.6704 | 0.6765 |
| Overall | - | - | - | 0.732 | 0.745 | 0.737 |

Table 1: The confusion matrix, and precision,recall and F-measure of the SemEval 2015 testset.

| | Training | Testing |
|---|---|---|
| **Questions** | 1300 | 200 |
| **Answers** | 6500 | 1001 |
| Direct | 1300 | 215 |
| Related | 1469 | 222 |
| Irrelevant | 3731 | 564 |

Table 2: Statistics of the training and testing data.

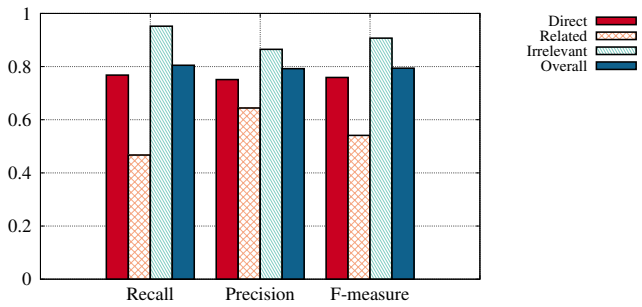


Figure 2: The training data cross validation results.

## 4 Evaluation

We evaluate our learning linguistic system by applying it on Fatwa questions/answers selection with a supervised learning framework.

### 4.1 Dataset Description

We train our classifier on the provided benchmark dataset in SemEval2015 (Màrquez et al., 2015). The used data is from Fatwa website [4]. Each question in the dataset is provided with five different answers. Each answer is labeled as Direct, Related, or Irrelevant. The distribution of the dataset we use is given in Table 2.

### 4.2 Results

In this section, we provide the experimental results of the training data and the SemEval 2015 test set.

Figure 2 shows the 10-folds cross validation results of the system training data using the two scores (the semantic relatedness and keyword matching scores). From the figure, the Direct and Irrelevant classes have better accuracies than the Related class. This is intuitive as the Related class is more general than the others (with few special marks), so it is more difficult to be classified.

[4]http://fatwa.islamweb.net/

Table 1 shows the confusion matrix of the SemEval 2015 test set results. The results also show that the Related class has lower accuracy than the Direct and Irrelevant. The overall system accuracy is $74.53\%$ and the system Macro-F1 is $67.65\%$

## 5 Conclusion

In this paper, we proposed our system to automate the process of Arabic answer selection in Community Question Answering systems where candidate answers are classified into answers that directly answer the question vs. those that can be helpful vs. those that are irrelevant. We constructed our knowledge-based system using a Quranic semantic ontology and the provided dataset in (Màrquez et al., 2015). The system first applies some preprocessing tasks over the question and answers, then a Semantic Interpreter converts the preprocessed sentences into weighted vectors of concepts. Using those vectors the system calculates a semantic score for each answer, which is fed, with an additional keyword matching score, into a decision tree classifier. The system has an overall accuracy of $74.53\%$.

## References


Abdul-Baquee M. Sharaf and Eric Atwell. 2012. *QurAna: Corpus of the Quran annotated with Pronominal Anaphora*. LREC.

Abdul-Baquee M. Sharaf and Eric Atwell. 2012. *QurSim: A corpus for evaluation of relatedness in short texts*. LREC.

Abu Bakr Al-Jaza'iri. 1986. *Aysar al-Tafasir li Kalaam il 'Aliyy il Kabir*.

Ahmed Magdy Ezzeldin, Mohamed Hamed Kholief, and Yasser El-Sonbaty. 2013. *ALQASIM: Arabic language question answer selection in machines*. In Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Springer, Berlin Heidelberg.

Eric Atwell, Claire Brierley, Kais Dukes, Majdi Sawalha and Abdul-Baquee Sharaf. 2011. *A An Artificial Intelligence Approach to Arabic and Islamic Content on the Internet*. Proceedings of NITS 3rd National Information Technology Symposium.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. *Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis*, volume 7. Proceedings of the 20th international joint conference on artificial intelligence.

Heba Abdelnasser, Reham Mohamed, Maha Ragab, Alaa Mohamed, Bassant Farouk, Nagwa El-Makky, and Marwan Torki. 2014. *Al-Bayan: An Arabic Question Answering System for the Holy Quran*. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP).

Hikmat Ullah Khan and Syed Muhammad Saqlain and Muhammad Shoaib and Muhammad Sher. 2013. *Ontology Based Semantic Search in Holy Quran.*, volume 2. International Journal of Future Computer and Communication, 570-575.

Ismail Ibn-Kathir. 1370. *Tafsir al-Qur'an al-Azim*.

Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2012. *IDRAAQ: New Arabic Question Answering System Based on Query Expansion and Passage Retrieval*. In CLEF.

Lluìs Màrquez and James Glass and Walid Magdy and Alessandro Moschitti and Preslav Nakov and Bilal Randeree. 2015. *SemEval-2015 Task 3: Answer Selection in Community Question Answering*. Proceedings of the 9th International Workshop on Semantic Evaluation.

Manal AlMaayah, Majdi Sawalha, and Mohammad AM Abushariah. 2014. *A Proposed Model for Quranic Arabic WordNet*. LRE-REL2, 9.

Nizar Habash, Owen Rambow and Ryan Roth. 2009. *MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization*. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Noorhan Hassan Abbas. 2009. *Quran'search for a Concept'Tool and Website*. M. Sc. thesis, University of Leeds (School of Computing).

Omar Trigui, Lamia Hadrich Belguith, Paolo Rosso, Hichem Ben Amor and Bilel Gafsaoui. 2012. *Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation*. CLEF (Online Working Notes/Labs/Workshop).

Saidah Saad, Naomie Salim, and Hakim Zainal. 2009. *Pattern extraction for Islamic concept.*, volume 2. Proceedings of IEEE 2nd. International Conference on Electrical Engineering & Informatics (ICEEI).

Saidah Saad, Naomie Salim, Hakim Zainal and S. Azman M. Noah. 2010. *A framework for Islamic knowledge via ontology representation.*. International Conference on Information Retrieval and Knowledge Management (CAMP).

Tim Buckwalter. 2002. *Arabic transliteration*. URL http://www.qamus.org/transliteration.htm.