CONSTRUCTING AN ONTOLOGY-BASED AND GRAPH-BASED KNOWLEDGE REPRESENTATION OF ENGLISH QURAN

Mohamad Fauzan Noordin\textsuperscript{a}, Tengku Mohd. Tengku. Sembok\textsuperscript{b}, Roslina Othman\textsuperscript{a}, and Ria Hari Gusmita\textsuperscript{a}

\textsuperscript{a}International Islamic University Malaysia, Malaysia
\textsuperscript{b}National Defence University of Malaysia, Malaysia

Abstract

This paper describes a work in constructing two models of knowledge representation (KR) in aiming to do evaluation of their achievement in contributing to increase performance of retrieving information on English Quran domain. Due to many approaches available to construct a KR in providing data for information retrieval process, there is a need to find out in what model the KR could provide a valuable contribution for retrieving information. We focused on ontology-based KR and graph database-based KR. We use Quranic Arabic corpus that available at http://www.corpus.quran.com as a source to build the KR. We extracted several data from it i.e. English token, token location, and token Part of Speech (POS). Protégé is used to construct the ontology and Neo4j is utilized in developing the graph database. Both KR models will be equipped in developing of an English Quran Question Answering system in order to evaluate their benefit.

Keywords: Knowledge representation (KR); English Quran; Ontology-based; Graph database-based

Abstrak


Kata Kunci: Perwakilan Pengetahuan (PP); Quran berbahasa Inggeris; Ontology-based; Graph database-based

©2016 Penerbit UTM Press. All rights reserved
1.0 INTRODUCTION

Knowledge Representation (KR) is the scientific domain concerned with the study of computational models able to explicitly represent knowledge by symbols and to process these symbols to produce new ones representing other pieces of knowledge work[1]. The knowledge can be gathered from a single person, an expert in a particular domain or from a well-defined document.

The Holy Quran, due to its unique style and allegorical nature, needs special attention about search and information retrieval issues[2]. There were many attempts in discovering an effective and efficient methods or architectures of retrieving information on the Holy Quran, ranging from corpus linguistics, knowledge representation, semantic interpretation, search engine, and question answering system. Especially for knowledge representation (KR), there are many approach can be applied to build it. Unfortunately, there is no sufficient information of in what model a KR can perform best in providing information for information retrieval application such as question answering system.

In this paper, we present a work of constructing two models of KR of English Quran i.e. ontology-based and graph-based KR aiming to discover in which form the KR can have a beneficial function in increasing question answering system performance.

In Section 2 we discuss existing works on knowledge representation construction, and in Section 3 we discuss our methodology to construct the KRs. Section 4 will be the end of the paper where it describes a conclusion.

2.0 RELATED WORKS

One of the works in employing ontology is [2], where there was the concept of ontology of semantic web that can be applied for carrying out a semantic search in Holy Quran. This work used a sample domain ontology that created based on living creatures including animals and birds mentioned in English Holy Quran. Several recommendation including model and framework containing creation of Quranic WordNet, integration, merging and mapping of domain ontologies under the umbrella of upper ontology were also presented.

Ontological work is getting widely used in many areas. An architecture of ontology-based domain specific natural language question answering system has been developed by [3]. This work proposed a step towards semantic web question answering (QA). The proposed architecture defines four basic modules suitable for enhancing existing QA capabilities with the ability of processing complex questions. Ontology and domain knowledge has a role in reformulating queries and identifying the relations. In 2014, the survey about existing ontology tools and methodologies were conducted in term of ontology’s support for constructing a knowledge representation [4].

An attempt to build a knowledge representation by using another approach i.e. knowledge graph came from [5]. This task proposed an architectural approach of representing knowledge graph for a complex question answering system. The knowledge graph was enriched by adding four kinds of entity relations consist of syntactic dependencies, semantic role labels, named entities, and co-reference links where has been proved effectively could be applied to answer a complex question.

3.0 RESEARCH METHODOLOGY

This part presents all methods to construct ontology-based and graph-based knowledge representation of English Quran.

3.1 Creating English Quran Corpus

In order to start construct the ontology and graph database, we need to have a collection of document that contain English Quran text called as English Quran corpus. Fortunately there exist a digital Quranic Arabic corpus that available at http://corpus.quran.com/ and it provides the English version of the corpus. Creating our corpus is done base on this following process as in Figure 2.

As depicted in Figure 1, construction of English Quran corpus utilizes several resources such as the existing Quranic Arabic Corpus. Quranic Arabic Corpus is an annotated linguistic resource that shows the Arabic grammar, syntax and morphology for each word in the Holy Quran. The corpus provides three levels of
analysis: morphological annotation, a syntactic treebank and a semantic ontology. This corpus is used as a resource to get each of Quranic term in English along with its part-of-speech. The display of Quranic Arabic Corpus is as in Figure 2.

Figure 2 shows that inside Quranic Arabic Corpus, there are 4 elements represented. These elements are as follows: Chapter, Verse, Token, and Character. All components represented by the corpus by using number with a format (chapter number: verse number: token number: character number). For example, (1:1:1:1) represents the first chapter in the Quran, the first verse, the first token, and the first character.

Quranic Arabic Corpus is used to construct English Quran corpus. As depicted in Figure 3, it can be seen that Quranic Arabic Corpus has characters part of each token. Unfortunately, there is no resource that can be used to map/translate those Quranic Arabic characters part into English term. This fact leads the English Quran corpus construction to only records 3 (three) elements i.e. chapter, verse, and token of Quranic domain. All token will be derived from Quranic Arabic Corpus along with its part-of-speech. Figure 3 shows display of Quranic Arabic Corpus that has a mapping to English term for every token inside.

Base on construction of English Quran corpus process that has been figured above, this is the steps to construct English Quran corpus:

2. Extract data resulted from point 1 to get English Quranic term, and it's part of speech
3. Write all result at Step 2 into a text file and XML file. XML file type is applied to support processing on the next steps such us creating graph database, ontology, and question answering system. XML file has a structure that follows standard form as described below:

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
  <token id="1">
    <location>(1:1:1)</location>
    <englishtranslation>In (the) name</englishtranslation>
    <listpos>
      <pos id="1" tag="P" description="prefixed preposition bi ">
        <pos id="2" tag="N" description="genitive masculine noun"/>
      </pos>
    </listpos>
  </token>
</corpus>
```

3.2 Construction of Ontology of English Quran

Development of English Quran ontology is conducted by using these following steps [6]:

1. Determine the domain and scope of the ontology. Domain of the ontology is English translation Quran. As the ontology will be used to assist natural language processing on question answering system, each concept derived from English translation Quran will have part of speech information.
2. Consider reusing existing ontologies. In developing the ontology, we consider reusing the existing English Quranic concepts ontology that has been built and presented at http://corpus.quran.com. All concepts defined in that existing ontology are used in this ontology development, and we enriched each of concept with its part of speech.
3. Define the classes and the class hierarchy. Since we decided to reuse the existing English Quranic ontology as mentioned at step 2, we also reused all classes and class hierarchy on it. This is sample of class hierarchy that we applied.
4. Define the properties of classes
Since an objective of ontology construction is to represent knowledge described in the Quran and provide verse(s) that refer it, each of classes on the ontology will have the same property i.e. resource that came from verse referring the class. In this research, we gave important property for each of class that will be used to increase the benefit of the ontology as a knowledge representation for any appropriate system. This important property is part of speech information that already exist as result of the first stage.

5. Create instances
We used location of verse that contain related concept as instance of that concept. Only classes at the lowest level that has instance.

We implemented the ontology construction by using a tool called as Protégé. Figure 4 displays result of ontology construction by using protégé:

In Figure 4, list of classes and the hierarchy as well is located at the left side. Instance of each class can be seen at the right side of the display.

3.3. Construction of Graph-based of English Quran

Construction of graph-based version of English Quran is conducted by using Neo4j and Java programming language with using Eclipse as the Integrated Development Environment (IDE). Since graph database applies a hierarchy, then the template of the corpus must be changed so that it accommodates data hierarchy i.e. Quran-Chapter-Verse-Token-POS. This modification is purposed to get a graph database from a given hierarchy. New format of English Quran corpus has this following XML format:

```
<quran>
  <chapter chaptternumber="1"
    chaptername="Al-Fatihih" totalverse="7">
    <verse versenumber="1">
      <token tokenid="1">
        <location>(1:1:1)</location>
        <englishtranslation>In (the)
          name</englishtranslation>
        <listpos>
          <pos id="1" tag="P"
            description="prefixed preposition bi "/>
          <pos id="2" tag="N"
            description="genitive masculine noun"/>
        </listpos>
      </token>
      ……………
    </verse>
  </chapter>
</quran>
```

Graph database that is created consists of 5 nodes where each of nodes has particular property with details shown in Table 1:

<table>
<thead>
<tr>
<th>Node Name</th>
<th>Property</th>
<th>Property Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALQURAN</td>
<td>AlQuranName</td>
<td>String “Al-Quran.”</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>ChapterName</td>
<td>Chapter name</td>
</tr>
<tr>
<td></td>
<td>ChapterNumber</td>
<td>Chapter number</td>
</tr>
<tr>
<td>VERSE</td>
<td>VerseNumber</td>
<td>Verse number</td>
</tr>
<tr>
<td>TOKEN</td>
<td>TokenWord</td>
<td>Verse token (word) in English</td>
</tr>
<tr>
<td></td>
<td>TokenID</td>
<td>Identity number of token</td>
</tr>
<tr>
<td>POS</td>
<td>PosId</td>
<td>Identity number of Part of Speech (POS)</td>
</tr>
<tr>
<td></td>
<td>PosTag</td>
<td>Value of POS</td>
</tr>
<tr>
<td></td>
<td>PosDescription</td>
<td>Description of POS value</td>
</tr>
</tbody>
</table>

The relationship between nodes is defined based on their original relationship on the Quran. As we know that Quran contains 114 chapters. Each chapter has particular number of verses. A verse is composed by several tokens. As the lowest level of object, each
token is completed by it’s part of speech information that represent token’s structure. All defined relationships are listed as follows:

Table 2 List of relationship between nodes on graph database

<table>
<thead>
<tr>
<th>Relationship Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTAINS</td>
<td>Relation between ALQURAN and CHAPTER</td>
</tr>
<tr>
<td>HAS</td>
<td>Relation between CHAPTER and VERSE</td>
</tr>
<tr>
<td>IS_COMPOSED</td>
<td>Relation between VERSE and TOKEN</td>
</tr>
<tr>
<td>HAS_STRUCTURE</td>
<td>Relation between TOKEN and POS</td>
</tr>
</tbody>
</table>

After defining all nodes and the relationship they have, we did construction of graph-based of English Quran. All steps that we applied are described below:

1. Declaration of Enumeration for Node Label
2. Declaration of Enumeration for Relationship
3. Initialize class object namely Graph Database Factory and Graph Database Service. These objects are used to create/read graph database. We also define a folder name to locate the graph database.
4. Initialize class object namely SAX Reader to read the corpus (namely as quran_reformat_v1.xml) and class object Document to store the content of the corpus into memory.
5. Initialize Transaction and call method begin Tx(). This is based on the concept that all database operation that access graph, index, or schema must be done in one transaction.
6. Declare doc Alquran to read corpus file and store it into memory
7. Get root element from corpus
8. Create a node from graph database for ALQURAN label and set it’s property
9. For each <chapter> tag at the corpus, conduct step 10 until 25
10. Get content of chapter name and chapter number attribute inside <chapter> tag
11. Create a node from graph database for Chapter label and set it’s property
12. Set a relationship between ALQURAN and CHAPTER
13. For each <verse> tag inside <chapter> tag at the corpus, conduct step 14 until 25 as follows
14. Get value of verse number attribute from <verse> tag
15. Create a node from graph database for label VERSE and set it’s property
16. Set relationship between CHAPTER and VERSE
17. For each <token> tag inside <verse> tag, conduct step 18 until 25
18. Get value of English translation attribute and tokenid from <token> tag
19. Create a node for TOKEN label and set it’s property
20. Set relationship between VERSE and TOKEN
21. For each <listpos> tag inside <token> tag, conduct step 22
22. For each of <pos> tag inside (listpos) tag, conduct step 23 until 25
23. Create a node for POS label and set it’s property
24. Set relationship between TOKEN and POS label
25. Save memory space and avoid out of memory as well
26. Record that the transaction is success
27. Close transaction
28. Shut down graph database

There are some kinds of query format that can be used to extract data from graph-based KR. Those query format and the result are discussed below:

1. Get a node with particular limit. For instance we want to get a node with limit 25, the query is: MATCH n RETURN n LIMIT 25
2. Get a node with label TOKEN and particular value. If we want to get a node with label token and value is Allah, this query format is applied:

   MATCH (token:TOKEN) WHERE token.TokenWord="Allah" return token

Result of the query is shown on Figure 5.

Figure 5 Get a node with limit 25
3. Get chapter, verse, token, and POS from token with particular value. As an example, it needs to have all nodes from token with value “the lord”, this following query is suitable:

```
MATCH (chapter:CHAPTER)-[:HAS]->(verse:VERSE)-[:IS_COMPOSED]->(token:TOKEN)-[:HAS_STRUCTURE]->(pos:POS) WHERE token.TokenWord = "the Lord" RETURN chapter, verse, token, pos
```

Result that is delivered is seen on Figure 7.

4. Get chapter, verse, token, and POS from token with particular value and there is only some token will be collected. If we need all nodes from token that has value “Ibrahim” and only 3 token is needed, then we can use this following query:

```
MATCH (chapter:CHAPTER)-[:HAS]->(verse:VERSE)-[:IS_COMPOSED]->(token:TOKEN)-[:HAS_STRUCTURE]->(pos:POS) WHERE token.TokenWord =~ "Ibrahim" RETURN chapter, verse, token, pos LIMIT 3
```

The query output nodes as figured out in Figure 8.
4.0 CONCLUSION

This paper has presented a construction process of two models of KR namely ontology and graph based KR. These two models are ready to be evaluated in order to discover which model perform best in providing information. The evaluation will be applied by developing an English Quran question answering system where each KR model will acts as a corpus. The best KR model is the one that question answering system deliver the highest number of correct answer from it. The authors hope there will be a useful information gained from the question answering system evaluation when using both of KR models.

References