Jurnal
Teknologi

Full Paper

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

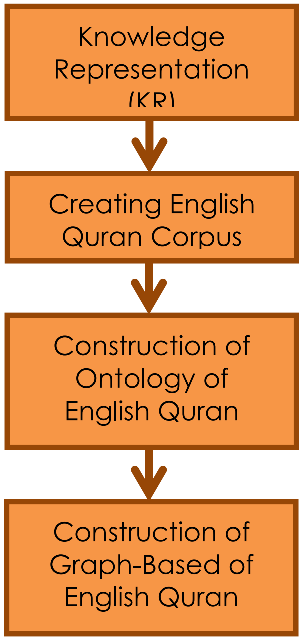
Article history
Received
14 November 2015
Received in revised form
20 March 2016
Accepted
2 March 2016

Mohamad Fauzan Noordin[a*], Tengku Mohd. Tengku. Sembok[b], Roslina Othman[a], and Ria Hari Gusmita[a]

[a]*International Islamic University Malaysia, Malaysia*
[b]*National Defence University of Malaysia, Malaysia*

*Corresponding author
fauzan@iium.edu.my

Graphical abstract


Knowledge Representation (KR)
Creating English Quran Corpus
Construction of Ontology of English Quran
Construction of Graph-Based of English Quran

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

Artkel ini membentangkan proses dalam mewujudkan dua model perwakilan pengetahuan (PP) yang bertujuan untuk menilai pencapaian mereka dalam menyumbang kepada peningkatan prestasi untuk memperolehi informasi Quran berbahasa Melayu. Oleh kerana terlalu banyak pendekatan untuk mewujudkan PP dalam menyediakan data untuk proses mendapatkan informasi, jenis model PP yang boleh menyumbang dalam proses memperolehi informasi adalah sangat diperlukan. Kami telah memberi tumpuan kepada dasar ontology dalam PP dan dasar graf pangkalan data dalam PP. Kami menggunakan Quranic Arabic Corpus yang boleh didapati di http://www.corpus.quran.com sebagai asas untuk membina PP. Kami telah mengekstrak beberapa data seperti token berbahasa Inggeris, token lokasi dan token Part of Speech (POS). Protégé telah digunakan untuk mewujudkan ontology dan Neo4j telah digunakan dalam membina graf pangkalan data. Kedua dua model PP akan dillengkapkan dalam mewujudkan sistem soalan jawapan Quran berbahasa Inggeris untuk menilai manfaatnya.

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

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.

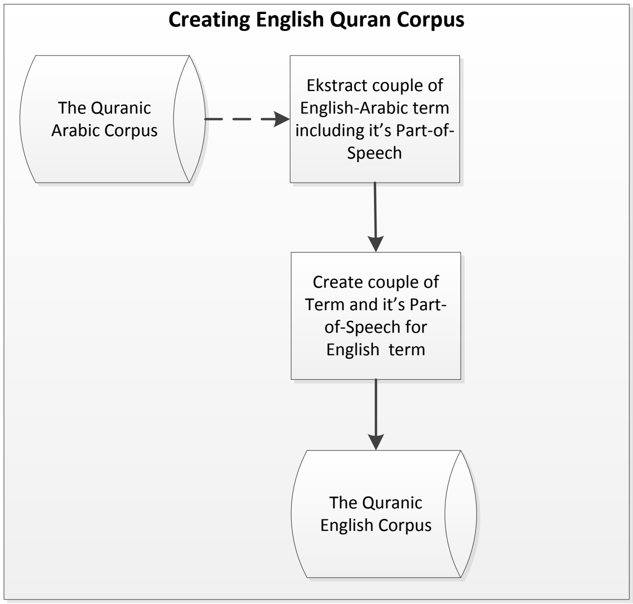
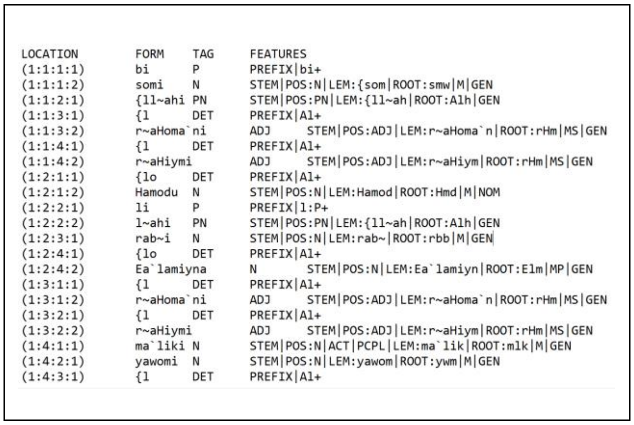

Creating English Quran Corpus
The Quranic Arabic Corpus
Ekstract couple of English-Arabic term including it's Part-of-Speech
Create couple of Term and it's Part-of-Speech for English term
The Quranic English Corpus

Figure 1. Creating English Quran corpus process

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.

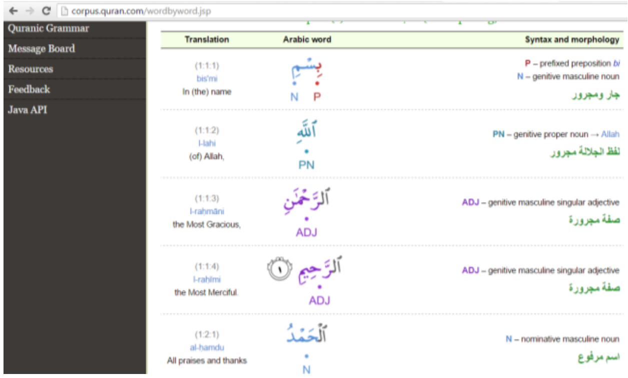
Figure 2 shows that inside Quranic Arabic Corpus, there are 4 elements represented. Those elements are as follows: Chapter, Verse, Token, and Character. All components represented by the corpus by using number with a format (chapter number: verse number: token number: character number). For example,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



```
LOCATION       FORM    TAG     FEATURES
(1:1:1:1)      bi      P       PREFIX|bi+
(1:1:1:2)      somi    N       STEM|POS:N|LEM:{som|ROOT:smw|M|GEN
(1:1:2:1)      {ll~ahi PN      STEM|POS:PN|LEM:{ll~ah|ROOT:Alh|GEN
(1:1:3:1)      {l      DET     PREFIX|Al+
(1:1:3:2)      r~aHoma`ni      ADJ     STEM|POS:ADJ|LEM:r~aHoma`n|ROOT:rHm|MS|GEN
(1:1:4:1)      {l      DET     PREFIX|Al+
(1:1:4:2)      r~aHiymi        ADJ     STEM|POS:ADJ|LEM:r~aHiym|ROOT:rHm|MS|GEN
(1:2:1:1)      {lo     DET     PREFIX|Al+
(1:2:1:2)      Hamodu  N       STEM|POS:N|LEM:Hamod|ROOT:Hmd|M|NOM
(1:2:2:1)      li      P       PREFIX|l:P+
(1:2:2:2)      l~ahi   PN      STEM|POS:PN|LEM:{ll~ah|ROOT:Alh|GEN
(1:2:3:1)      rab~i   N       STEM|POS:N|LEM:rab~|ROOT:rbb|M|GEN
(1:2:4:1)      {lo     DET     PREFIX|Al+
(1:2:4:2)      Ea`lamiyna      N       STEM|POS:N|LEM:Ea`lamiyn|ROOT:Elm|MP|GEN
(1:3:1:1)      {l      DET     PREFIX|Al+
(1:3:1:2)      r~aHoma`ni      ADJ     STEM|POS:ADJ|LEM:r~aHoma`n|ROOT:rHm|MS|GEN
(1:3:2:1)      {l      DET     PREFIX|Al+
(1:3:2:2)      r~aHiymi        ADJ     STEM|POS:ADJ|LEM:r~aHiym|ROOT:rHm|MS|GEN
(1:4:1:1)      ma`liki N       STEM|POS:N|ACT|PCPL|LEM:ma`lik|ROOT:mlk|M|GEN
(1:4:2:1)      yawomi  N       STEM|POS:N|LEM:yawom|ROOT:ywm|M|GEN
(1:4:3:1)      {l      DET     PREFIX|Al+
```

Figure 2. Display of Quranic Arabic Corpus

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.



Translation	Arabic word	Syntax and morphology
(1:1:1) bis'mi In (the) name	بِسْمِ N P	P – prefixed preposition *bi* N – genitive masculine noun جار ومجرور
(1:1:2) l-lahi (of) Allah,	ٱللَّهِ PN	PN – genitive proper noun → Allah لفظ الجلالة مجرور
(1:1:3) l-raḥmāni the Most Gracious,	ٱلرَّحْمَٰنِ ADJ	ADJ – genitive masculine singular adjective صفة مجرورة
(1:1:4) l-raḥīmi the Most Merciful.	ٱلرَّحِيمِ ١ ADJ	ADJ – genitive masculine singular adjective صفة مجرورة
(1:2:1) al-ḥamdu All praises and thanks	ٱلْحَمْدُ N	N – nominative masculine noun اسم مرفوع

Figure 3 Display of Word-by-Word Menu in Quranic Arabic Quran Website

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

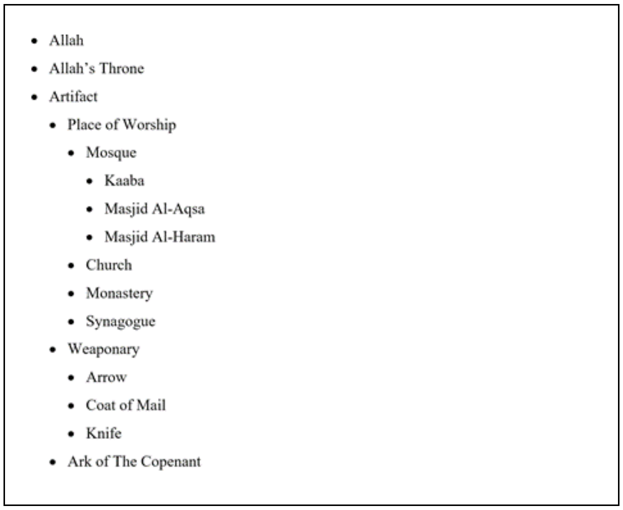
1. Get data of a couple of English-Arabic Quranic term including its part of speech from http://www.corpus.quran.com/.
2. Extract data resulted from point 1 to get English Quranic term, and it's part of speech
3. Write all result at Step 2 into a text file and XML file. XML file type is applied to support 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"/>
   </listpos>
 </token>
<corpus>
```

3.2 Construction of Ontology of English Quran

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

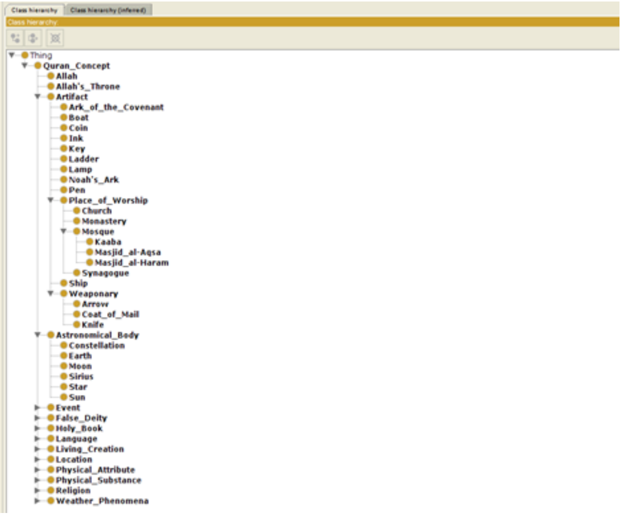
1. Determine the domain and scope of the ontology. Domain of the ontology is English translation Quran. As the ontology will be used to assist natural language processing on question answering system, each concept derived from English translation Quran will have part of speech information.
2. Consider 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.



- Allah
- Allah's Throne
- Artifact
  - Place of Worship
    - Mosque
      - Kaaba
      - Masjid Al-Aqsa
      - Masjid Al-Haram
    - Church
    - Monastery
    - Synagogue
  - Weaponary
    - Arrow
    - Coat of Mail
    - Knife
  - Ark of The Copenant

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 come 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:



- Thing
  - Quran_Concept
    - Allah
    - Allah's_Throne
    - Artifact
      - Ark_of_the_Covenant
      - Boat
      - Coin
      - Ink
      - Key
      - Ladder
      - Lamp
      - Noah's_Ark
      - Pen
      - Place_of_Worship
        - Church
        - Monastery
        - Mosque
          - Kaaba
          - Masjid_al-Aqsa
          - Masjid_al-Haram
        - Synagogue
      - Ship
      - Weaponary
        - Arrow
        - Coat_of_Mail
        - Knife
    - Astronomical_Body
      - Constellation
      - Earth
      - Moon
      - Sirius
      - Star
      - Sun
    - Event
    - False_Deity
    - Holy_Book
    - Language
    - Living_Creation
    - Location
    - Physical_Attribute
    - Physical_Substance
    - Religion
    - Weather_Phenomena

Figure 4 Result of English Quran ontology construction

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 chapternumber="1"
chaptername="Al-Fatihah" 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 1 List of node and property value on graph database

Node Name	Property	Property Value
ALQURAN	AlQuranName	String "Al-Quran."
CHAPTER	ChapterName	Chapter name
	ChapterNumber	Chapter number
VERSE	VerseNumber	Verse number
TOKEN	TokenWord	Verse token (word) in English
	TokenID	Identity number of token
POS	PosId	Identity number of Part of Speech (POS)
	PosTag	Value of POS
	PosDescription	Description of POS value

The relationship between nodes is defined base on their original relationship on the Quran. As we know that Quran contains 114 chapter. Each chapter has particular number of verses. A verse is composed by several token. 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

Relationship Name	Description
CONTAINS	Relation between ALQURAN and CHAPTER
HAS	Relation between CHAPTER and VERSE
IS_COMPOSED	Relation between VERSE and TOKEN
HAS_STRUCTURE	Relation between TOKEN and POS

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 dow m 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
   ```
   Result of the query is shown on Figure 5.

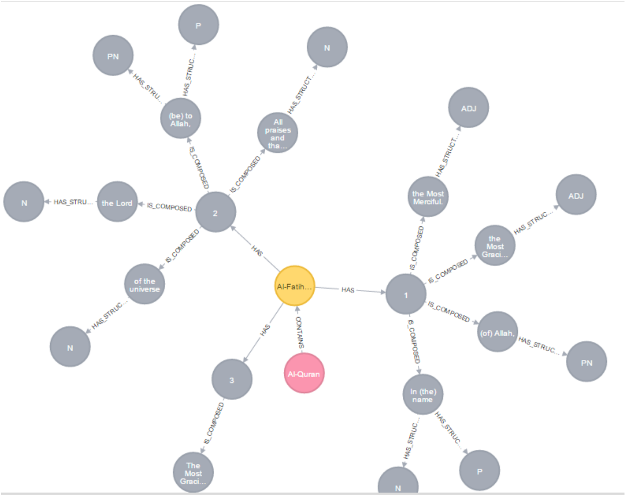
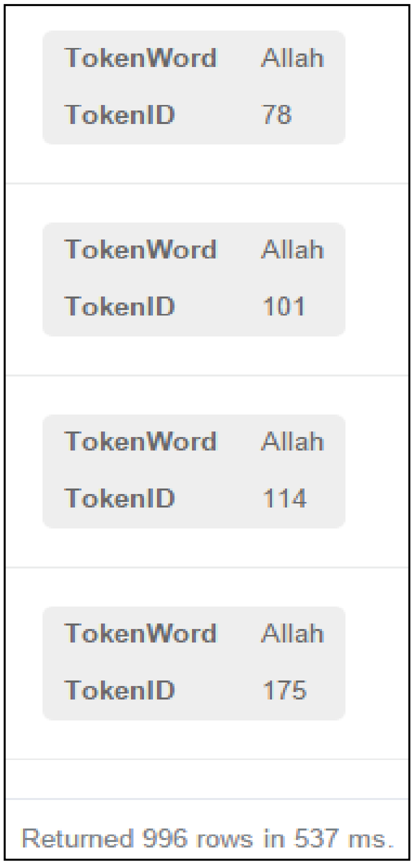


Figure 5 Get a node with limit 25

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 depicted by Figure 6.



```
TokenWord   Allah
TokenID     78

TokenWord   Allah
TokenID     101

TokenWord   Allah
TokenID     114

TokenWord   Allah
TokenID     175

Returned 996 rows in 537 ms.
```

Figure 6 Get a node with label TOKEN that has Token Word=Allah

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.

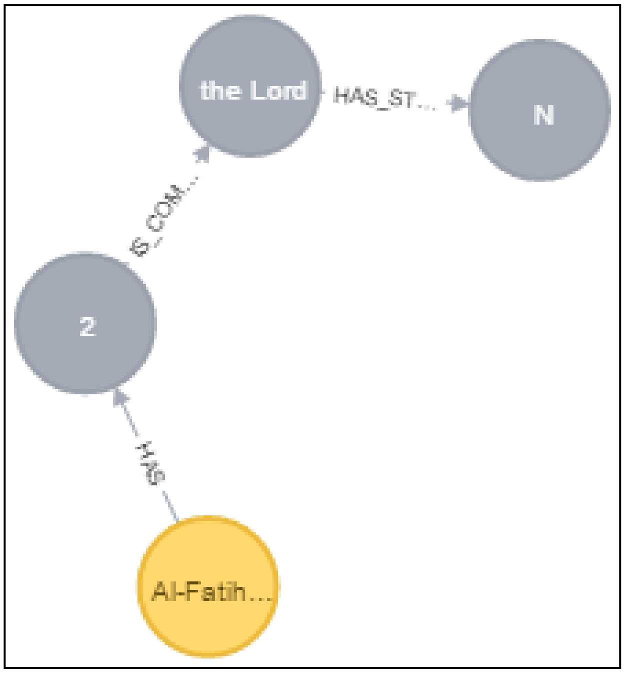


Figure 7 Get chapter, verse, token, and POS from token=the Lord

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.

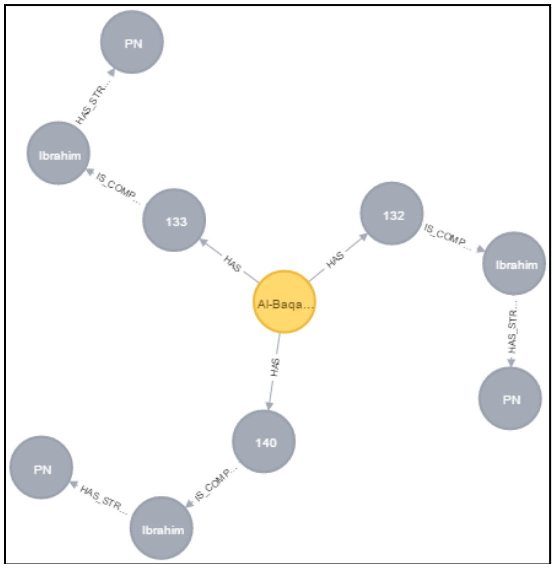


Figure 8 Get chapter, verse, token, and POS from token that valued Ibrahim and only three tokens that will be collected

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

[1] M. Chein, M. Mugnier. 2009. Graph-based Knowledge Representation, London, *Springer Verlag*,

[2] H. U. Khan, S. M. Saqlain, et al. 2013. Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*. 2(6).

[3] P.M. Athira et al. 2013. Architecture of an Ontology-Based Domain Specific Natural Language Question Answering System. *International Journal of Web and Semantic Technology*. 4(4).

[4] S. Jain and S. Mishra. 2014. Knowledge Representation with Ontology Tools and Methodology. *International Journal of Computer Applications*.

[5] T. Jurczyk and J. D. Choi. 2015. Semantic-based Graph Approach to Complex Question-Answering. *Proc of NAACL-HLT 2015 Student Research Workshop (SRW)*. 140-146.

[6] N. F. Noy and D. L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford University*. [Online]. Available: http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

[7] T. M. T. Sembok. 2015. Knowledge Representation in Information Retrieval. *Journal of Recent Advances in Computer Science*.

[8] D. H. Deshmukh and S. D. Deshpande. 2013. A Review of Ontology based Information Retrieval. *International Journal of Advance Research in Computer Science and Management Studies.* 1.

[9] Yin, Wenke, Weiyi Ge, and Heng Wang. 2014. CDQA: An Ontology-Based Question Answering System For Chinese Delicacy. Cloud Computing and Intelligence Systems (CCIS). *2014 IEEE 3rd International Conference on. IEEE*, 2014.

[10] Shvaiko, Pavel, and Jérôme Euzenat. 2013. Ontology Matching: State Of The Art And Future Challenges. Knowledge and Data Engineering, IEEE Transactions on 25(1): 158-176.

[11] Zou, Lei, et al. 2014. Natural Language Question Answering Over RDF: A Graph Data Driven Approach." *Proceedings of the 2014 ACM SIGMOD International Conference On Management Of Data*. ACM.

[12] West, Robert, et al. 2014. Knowledge Base Completion Via Search-Based Question Answering. *Proceedings Of The 23rd International Conference On World Wide Web*. ACM.

[13] Lopez, Vanessa, et al. 2013. Evaluating Question Answering Over Linked Data. *Web Semantics: Science, Services and Agents on the World Wide Web*. 21: 3-13.

[14] Kamsu-Foguem, Bernard, Gayo Diallo, and Clovis Foguem. 2013. Conceptual Graph-Based Knowledge Representation For Supporting Reasoning In African Traditional Medicine. *Engineering Applications of Artificial Intelligence* 26(4): 1348-1365.

[15] Angles, Renzo. 2012. A Comparison Of Current Graph Database Models. *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on*. IEEE.