Proceedings Paper:
AlMaayah, M, Sawalha, M and Abushariah, M (2014) A proposed model for Quranic Arabic WordNet. In: Brierley, C, Sawalha, M and Atwell, E, (eds.) Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts, 31 May 2014, Reykjavik, Iceland. LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop, 31 May 2014, Reykjavik, Iceland. LRA , 9 - 13.

A Proposed Model for Quranic Arabic WordNet

Manal AlMaayah[1], Majdi Sawalha[2], Mohammad A. M. Abushariah[3]
Computer Information Systems Department, King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan.
E-mail: [1]manalmaayah@ymail.com, [2]sawalha.majdi@gmail.com, [3]m.abushariah@ju.edu.jo

Abstract

Most recent Arabic language computing research focuses on modern standard Arabic, but the classical Arabic of the Qur'an has been relatively unexplored, despite the importance of the Qur'an to Islam worldwide which can be used by both scholars and learners.This research work proposes to develop a WordNet for Qur'an by building semantic connections between words in order to achieve a better understanding of the meanings of the Qur'anic words using traditional Arabic dictionaries and a Qur'an ontology. The Qur'an corpus will be used as text and Boundary Annotated Qur'an Corpus (Brierley et al, 2012) will be used to explore the root and Part-of-Speech for each word and the word by word English translations. Traditional Arabic dictionaries will be used to find the Arabic meaning and derived words for each root in the Qur'anic Corpus. Then, these words and their meanings (Arabic, English) will be connected together through semantic relations. The achieved Qur'anic WordNet will provide an integrated semantic Qur'anic dictionary for the Arabic and English versions of the Qur'an.

Keywords: Qur'an WordNet, Qur'an ontology, Arabic dictionaries.

1. Introduction

The Holy Qur'an is the central religious text of Islam. The Qur'an is considered to be an excellent gold standard text that is essential for developing, modeling and evaluating Arabic NLP tools. The Qur'an as a corpus and is made up of 77,430 words. It is divided into 114 chapters which consist of 6,243 verses. The Qur'anic WordNet services anyone who seeks to expand his knowledge of Qur'anic Arabic vocabulary and increases understanding of the Qur'an and of Islam. In Qur'an, we find many words that are conceptually synonyms but if we investigate their dictionary meanings, then differences will surface. For example:أحمد*ahmad* (SAW), المزّمّل*al-muzzammil*, المدّثّر*al-muddaṯṯir*, يس *yāsīn*, الرّسول*ar-rasūl*are synonymous of Muhammad (SAW). Another example is سَبِيلِ *sabīl* and وَجْهِ *waǧh* are synonyms of (the way that they spend their wealth). Table 1 illustrates this example.

Table 1: Examples of سَبِيلِ sabīl and وَجْهِ waǧh in different verses.

Example 1: Chapter 2, verse 272
وَمَا تُنْفِقُونَ إِلَّا ابْتِغَاءَ **وَجْهِ** اللَّهِ
and you do not spend except seeking the countenance of Allah
Example 2: Chapter 2, verse 261
مَثَلُ الَّذِينَ يُنْفِقُونَ أَمْوَالَهُمْ فِي **سَبِيلِ** اللَّهِ كَمَثَلِ حَبَّةٍ أَنْبَتَتْ سَبْعَ سَنَابِلَ
The example of those who spend their wealth in the way of Allah is like a seed [of grain] which grows seven spikes

Therefore, we will get a better understanding of the meanings of the Qur'an by modeling a Qur'anic WordNet and developing a computational linguistic theory for Arabic using new technologies of NLP, traditional Arabic linguistic theory and classical Arabic dictionaries. We can utilize Qur'anic WordNet for machine learning tasks, such as Word Sense Disambiguation (WSD) where the word may have many meanings and it becomes therefore crucial to distinct the different senses. For example, the word وجه *waǧh* has three senses as illustrated in Table (2). To decide the returned sense in a current context, the Qur'anic WordNet is essential.

Table 2: Examples of different senses of the word وجه *waǧh*

لِلَّهِ وَمَنِ اتَّبَعَنِ	وَجْهِيَ	فَإِنْ حَاجُّوكَ فَقُلْ أَسْلَمْتُ
Submitted myself to Allah		
النَّهَارِ	وَجْهَ	آمِنُوا بِالَّذِي أُنزِلَ عَلَى الَّذِينَ آمَنُوا
At the beginning of the day		
	وَجْهِهَا	ذَٰلِكَ أَدْنَىٰ أَن يَأْتُوا بِالشَّهَادَةِ عَلَىٰ
It's true form		

This paper is structured as follows: section 2 a brief overview of the Qur'anic WordNet, section 3 related work, section 4 methodology, section 5 conclusion and future work.

2. A Brief Overview of the Qur'anic WordNet

The Qur'anic WordNet is a multidisciplinary project that contributes to Information Technology including; Computational Linguistics and Language Engineering, Information Extraction, Text Analytics, Text Data Mining, and Machine Learning, and to other disciplines such as Islamic Studies, Linguistics, Arabic Linguistics and Lexicography.

Qur'anic WordNet will make use of Arabic WordNet (Elkateb et al, 2006), Qur'an Ontology and classical

Arabic dictionaries. We will provide a literature investigation on WordNets, Arabic WordNet and Qur'an WordNet and ontologies. Then we will design a method for measuring semantic similarity between words included in the Qur'anic WordNet.

3. Related Work

Several research initiatives were directed towards building a WordNet for various foreign languages. WordNet was first developed for English in 1980s at the Cognitive Science Laboratory of Princeton University, hence is known as Princeton WordNet. This is a large-scale lexical database that was manually constructed (Miller,1995), (Miller and Fellbaum, 2007), (Fellbaum and Vossen, 2012).

George A. Miller (1995) showed the importance of WordNet for the English language and he outlined the semantic relations that provide more effective combinations of traditional lexicographic information and modern computing. English nouns, verbs, adjectives, and adverbs are organized into sets of synonyms, each representing a lexicalized concept, and semantic relations which link the sets of synonym. He used more than 116.000 pointers between words and word senses to build these relations. The semantic relations that were used in WordNet include (synonymy, antonymy, hyponymy, meronymy, troponomy, entailment), in addition to the senses. The interface of WordNet is easy to use with all word forms. When a word is entered, its syntactic category appears in menu. Once the appropriate syntactic category is selected, semantic relations for that word are displayed.

In 1998, the EuroWordNet of eight European languages was developed and linked to the English WordNet. This is a resource that provides multilingual lexical database. EuroWordNet contributed to variety of fundamental innovations in the design of the WordNet. It defines a set of base concepts and increases the connectivity among synsets (Fellbaum and Vossen, 2012).

In 2000, the idea of the Global WordNet was founded by Fellbaum and Vossen (Fellbaum and Vossen, 2007) and (Vossen and Fellbuam, 2012) for the aim of establishing a WordNet that supports a large number of languages interlinked into a single knowledgebase to facilitate inter communicability. Currently, Global WordNet covers sixty distinct languages, including: Arabic, Bantu, Basque, Chinese, Bulgarian, Estonian, Hebrew, Icelandic, Japanese, Kannada, Korean, Latvian, Nepali, Persian, Romanian, Sanskrit, Tamil, Thai, Turkish, Zulu, etc.

Elkateb and Black (2006) introduced the project for building a lexical resource for Modern Standard Arabic based on the English language's Princeton WordNet (Fellbaum, 1998) and the standard word representation of senses. The tool that was built has a lexicographer's interface and can be linked directly to EuroWordNet (EWN).

Trad and Koroni (2012) used Arabic WordNet Ontology for query expansion. They evaluated how beneficial was it compared to the English WordNet Ontology. Two individual corpora were used, one for Arabic documents and the other for English documents. Scanning and indexing files and documents are made via a multithreaded procedure to maintain the best interactivity and efficiency. The result of the comparison showed that the English ontology was better and more global than the Arabic ontology.

Shoaib (2009) proposed a model that is capable of performing semantic search in the Qur'an. The model exploits WordNet relationships in a relational database model. The implementation of this model has been carried out with the latest technologies and Surah Al-Baqrah (Chapter 2 from the Qur'an) has been taken as a sample text. The model facilitates performing a subject search for Qur'an readers and provides a framework capable of retrieving related verses from the Qur'an whether the query keyword is present in them or not. Semantic search is carried out in two steps. The first step is to identify only one sense of the query keyword using (WSD). The second step is to retrieve all synonyms of the identified sense of the word. The main goal of this work is to improve the semantic search and retrieval over Qur'anic topics.

4. Our proposed model for the Arabic Qur'anic WordNet

The proposed work "Arabic Qur'anic WordNet" will be implemented and evaluated (as illustrated in figure 1) in the following steps:

- **Qur'an text is preprocessed** using (i) tokenization, (ii) elimination of stop words, (iii) stemming and (iv) Part-Of-Speech tagging for each word in the Qur'an corpus.

- **Synsets (synonym sets)** are generated by grouping words of similar meaning and part-of-speech. For example, the words {رأى *rā* , أبصَر *abaṣara*, نظَر *nazara*} that share the sense, "see", are grouped together in a synset.

- **Semantic relations between different synsets are defined**. The semantic relationships that will be included in the Qur'anic WordNet are:

  a. Synonymy is determined. The words that have similar meanings are synonyms. For example, the words {حول *ḥawl* , سنه *sanah*, عام *ʿām*} are synonyms, they all mean "a year".

  b. Antonyms are marked. The words that have opposite meanings such as الحياة *al- ḥayāt* "life", and الموت *al-mawt* "death" are labeled 'antonyms'.

  c. A Glossary is compiled. This is used to store the glosses for every synset. Gloss may contain an

explanation, definition, and example of sentences. For instance, {المطر *al-maṭar*, الغيث *al-ġayṯ*} is a synset that share the sense "rain". However, they are used in different contexts (see Table 3).

d. **Similarity** between concepts is differentiated by connecting synsets that have similar meanings. For example, {خشيه *khāshīnyh*, خوف *khawf*} and {الروع *ar-rawʿu*, الرهب *ar-rahb*} is a synset that share the meaning of "fear" or "fright", see Table 4.

The implementation of the Qur'anic WordNet utilizes the Qur'an corpus as text and the Boundary Annotated Qur'an Corpus (Brierley et al, 2012) for exploring roots, POS tags, and English meaning for each word in the corpus. After that, the Arabic meaning and the derived words for each root are found using classical Arabic dictionaries. Finally, these words and their meanings (Arabic, English) are connected together with semantic relations.

The Qur'anic WordNet will be evaluated using a suitable evaluation technologies, standards, and metrics. We will build a gold standard for evaluating Qur'anic WordNet which at the same time can be used for evaluating Arabic WordNet.

Table 3: Glossary Example

Word	Semantics	Example	Translation
al-maṭar المطر	torment	وَأَمْطَرْنَا عَلَيْهِمْ **مَطَرًا** فَانْظُرْ كَيْفَ كَانَ عَاقِبَةُ الْمُجْرِمِينَ (الأعراف: 84)	And We rained upon them a rain [of stones]. Then see how was the end of the criminals.
al-ġayṯ الغيث	goodness and grace	وَهُوَ الَّذِي يُنَزِّلُ **الْغَيْثَ** مِنْ بَعْدِ مَا قَنَطُوا وَيَنْشُرُ رَحْمَتَهُ ۚ وَهُوَ الْوَلِيُّ الْحَمِيدُ (الشورى: 28)	And it is He who sends down the rain after they had despaired and spreads His mercy. And He is the Protector, the Praiseworthy.

Table 4.a: Example of similar meaning{خشيه khāshīnyā, خوف khawf}

Word	Semantics	Example	Translation
خشيه *khāshīnyā*	as (they) fear	إِذَا فَرِيقٌ مِنْهُمْ يَخْشَوْنَ النَّاسَ **كَخَشْيَةِ** اللَّهِ (النساء: 77)	at once a party of them feared men as they fear Allah or with [even] greater fear.
خوف *khawf*	fear	فَمَا آمَنَ لِمُوسَىٰ إِلَّا ذُرِّيَّةٌ مِنْ قَوْمِهِ عَلَىٰ **خَوْفٍ** مِنْ فِرْعَوْنَ وَمَلَئِهِمْ أَن يَفْتِنَهُم (يونس: 83)	But no one believed Moses, except [some] youths among his people, for fear of Pharaoh and his establishment that they would persecute them.

Table 4.b:Example of similar meaning{الروع *ar-rawʿu*, الرهب *ar-rahb*}

Word	Semantics	Example	Translation
الروع ***l-rawʿu***	fright	فَلَمَّا ذَهَبَ عَنْ إِبْرَاهِيمَ **الرَّوْعُ** وَجَاءَتْهُ الْبُشْرَىٰ يُجَادِلُنَا فِي قَوْمِ لُوطٍ (هود:74)	And when the fright had left Abraham and the good tidings had reached him, he began to argue with Us concerning the people of Lot.
الرهب ***l-rahb***	fear	وَاضْمُمْ إِلَيْكَ جَنَاحَكَ مِنَ **الرَّهْبِ** (القصص:32)	And draw in your arm close to you [as prevention] from fear

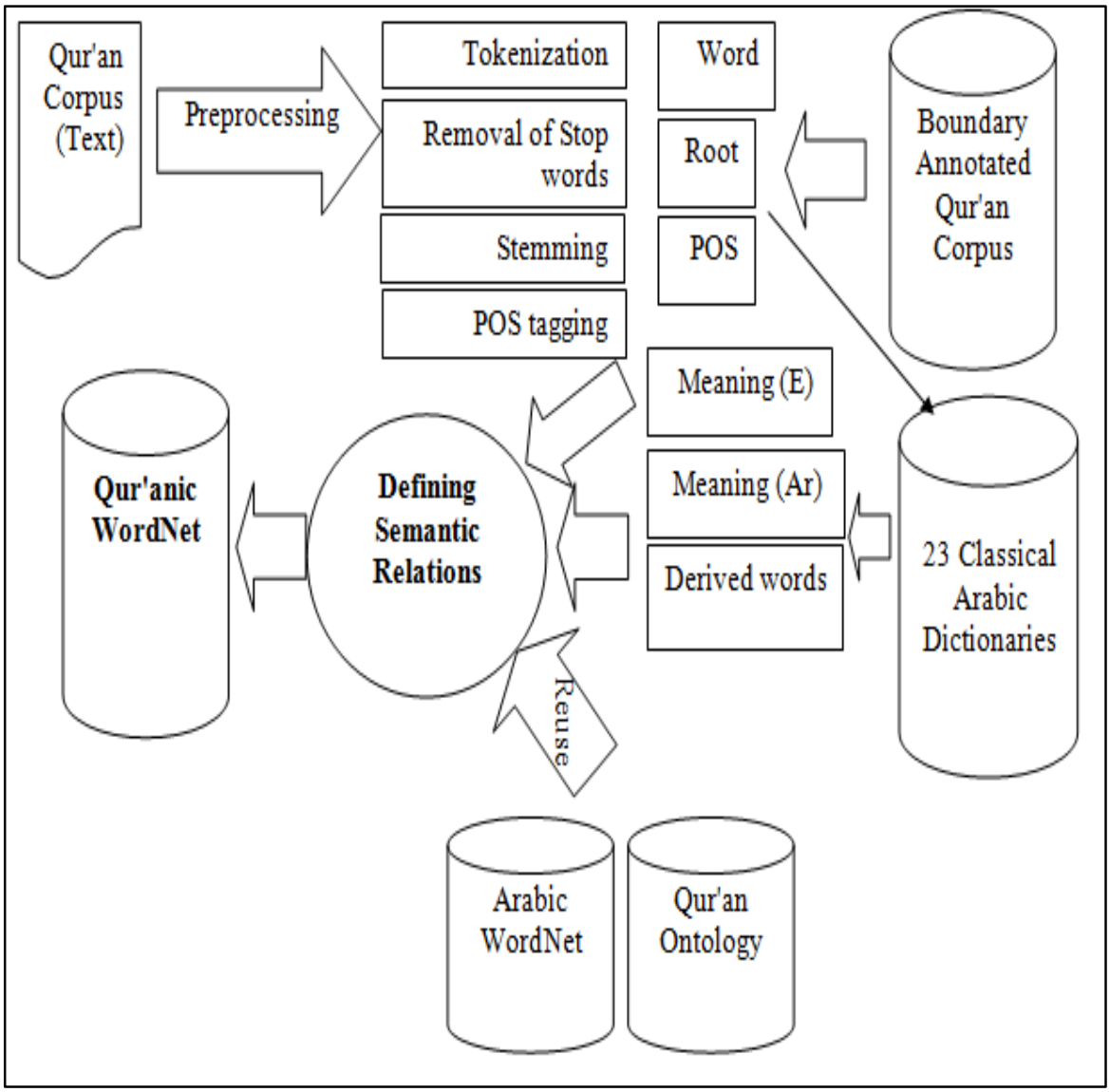

Figure 1: Implementation and evaluations model for building the Qur'anic WordNet

5. Conclusion

This work proposes the Arabic Qur'anic WordNet as a valuable resource designed and built for the study the semantic relations between Arabic concepts in the Qur'an. This work is unprecedented. The proposed model will be implemented on Qur'anic words using Arabic WordNet, classical Arabic dictionaries, and Qur'an ontology. Suitable evaluation methods will be designed, implemented and applied. These shall also be used as a standard for evaluating Arabic WordNet. We will design a methodology for measuring semantic similarity between different words included in the Qur'anic WordNet using edge-counting techniques. After building the Quranic WordNet, we plan to facilitate it for machine learning tasks such as word sense disambiguation.

References

Black, W., Elkateb, S., & Vossen, P. (2006). Introducing the Arabic wordnet project. In *Proceedings of the third International WordNet Conference (GWC-06).*

Brierley, C., Sawalha, M. Atwell E. (2012). open-source boundary-annotated corpus for Arabic speech and language processing. *LREC 2012, Istanbul, Turkey*.

Dukes, K. (2012). The Qur'anic Arabic Corpus. Online. *http://corpus.quran.com* .

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). Building a wordnet for arabic. In *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006).*

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Fellbaum, c. Vossen P. 2012. Challenges for a multilingual WordNet. *Lang Resources & Evaluation*, 46, 313–326.

Fellbaum, C., Vossen, P. (2007). Connecting the universal to the specific. In T. Ishida, S. R. Fussell & P. T. J. M. Vossen (Eds.), Intercultural collaboration: *First international workshop* (Vol. 4568, pp. 1–16). *Lecture Notes in Computer Science*, Springer, New York.

Khan, H., Saqlain, S.M., Shoaib, M., & Sher, M. (2013). Ontology Based Semantic Search in Holy Quran. *International Journal of Future Computer and Communication*. 2, 560-575.

M. Shoaib, M. N. Yasin, U. K. Hikmat, M. I. Saeed, & M. S. H.Khiyal.(2009). Relational Word Net model for semantic search in holy quran. In *Proceedings of International Conference on Emerging Technologies, Pakistan.* pp. 29-34.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*. 38, 39–41.

Miller G. A., Fellbaum C. (2007). WordNet then and now. *Lang Resources & Evaluation.* 41, 209–214.

Muhammad, A. B. (2012). Annotation of conceptual co-reference and text mining the Qur'an. *University of Leeds.*

Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Antonia Martí, M., Black, W., Elkateb, S., Kirk,J., Pease, A., Vossen, P.,& Fellbaum, C. (2008). Arabic WordNet: Current State and Future Extensions .In *Proceedings of the Fourth International GlobalWordNet Conference-GWC 2008. Szeged, Hungary.*

Sharaf, A., Atwell, E. (2012). QurSim: A corpus for evaluation of relatedness in short texts. *LREC 2012, Istanbul, Turkey.*

Trad, R., Mustafa, H., Koroni, R., & Almaghrabi, A. (2012). Evaluating Arabic WordNet Ontology by expansion of Arabic queries using various retrieval models. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on* (pp. 155-162). IEEE.

Vossen, P. (2002). EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document.