Bulletin of Electrical Engineering and Informatics
Vol. 4, No. 3, September 2015, pp. 241~247
ISSN: 2089-3191

Word Sense Disambiguation on English Translation of Holy Quran

Sarah Abdul-Ameer Mussa, Sabrina Tiun
Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia
Jalan Reko, 43600 Bangi, Malaysia, Selangor
e-mail: rosa_sara90@yahoo.com

Abstract

This article proposes a system based on the interpretation on the Quranic text that has been translated into English language using word sense disambiguation. This system is based on a combination of three traditional semantic similarity measurements, which are Wu-Palmer (WUP), Lin (LIN), and Jiang-Conrath (JCN) for word sense disambiguation on the English Al-Quran. The experiment was performed to obtain the best overall similarity score. The empirical results demonstrate that the combination of the three mentioned semantic similarity techniques obtained competitive results when compared with using individual similarity measurements.

Keywords*: Quranic IR, Semantic similarity, Quranic Translated Text, WordNet*

1. Introduction

The Holy Quran is the central religious verbal text of Islam, and the right understanding of the Quranic text is very necessary for Muslim people. It is the religious text of more than 1.5 billion Muslims around the world, who speak in different languages. It was consists of 114 ‘surah’ (chapters), which have obvious textual boundaries. In general, the longer ‘surah’ appear earlier in the Quran, while the shorter ones appear later. Each ‘surah’ comprises several ‘ayat’ or verses. Neither the number of verses in the ‘surah’ nor the word count of the verses is the same. The Arabic Quranic corpus consists of 77,784 word tokens and 19,287 word types [14, 15, 16].

Although Muslims read the Arabic text of the Holy Quran, it can be helpful to have its translation and well-formed interpretation in the mother tongue of every nation or in an internationally studied language like English in order to create a better understanding of the Quran. It was originally written in Arabic and therefore when it is translated into other languages the closest meaning among various possible choices presents an innate challenge. When translations are carried out, there is always a degree of human judgment whereby the translator endeavours to select the best interpretation. Even though modern linguistics brings clarity and understanding, there may still be a lingering doubt about whether the original meaning is being conveyed. Disambiguation identifies words that have more than one meaning so there is no ambiguity. Often this is clear from the context of the concepts being communicated. The process of defining meaning is also relevant to computer-related writing, including internet search engines. Writing can contain implied meaning, for example by the use of inference or reciprocal pronouns, which are interpreted by the reader as part of coherent understanding. In computational linguistics, word sense disambiguation (WSD) is a technique that resolves ambiguity by analysing the context in which they are written. For example, the concepts associated with the word ‘issue’ include ‘giving an item to a person’, ‘a particularly copy of a publication’, or ‘a difficulty that needs overcoming’. The WSD concept is an integral and complex part of natural language processing. The complexity has to be resolved by other methods than human interpretation. The process must overcome ambiguity by identifying the intended sense, including by algorithms that evaluate language. The style in which the verse of the Quran are written poses a challenge for humanity to dispel any confusion and grasp the intended meaning, as some words and phrases are ambiguous as the component words convey various senses or are polysemous. Problems arise in word sense disambiguation in relation to words that do not have a finite meaning and when the sense requires interpretation. To resolve

Received February 3, 2015; Revised May 1, 2015; Accepted May 25, 2015

the ambiguity a forced choice has to be made that establishes the closest fit of the meaning to the word. There has been extensive research to find the best approach and method for word sense disambiguation, carried out in various languages.

Quranic text Information Retrieval (IR) is quite demanding yet very trivial. As such, users will not always use the exact keywords to retrieve the relevant Quranic text (verse). Many have tried to overcome this problem by expanding or reformulating the query entered by users by using semantic approaches and resources such as ontologies and thesauri. Word Sense Disambiguation (WSD) has been less interesting to the IR research community due to its insignificant or very little significant impact on the IR performance. Recently, researchers have been interested in applying WSD to the IR problem, believing that an in-depth semantic analysis of the query process will have a good impact on the IR performance. However, there have not been any studies so far that mention the use of WSD for Quranic IR. As such, it is assumed that very little or no research on WSD for Quranic IR has been carried out. Thus, this research is motivated to create a WSD that somehow can be used to enhance any Quranic domain application. This perceived gap motivated the direction of this research, which examines its performance in this context. The concepts of similarity or relatedness are central to natural language processing functions such as word sense disambiguation, machine-based translation, analysis of discourse structure, classifying, summarising and annotating text, information extraction and retrieval, automated indexing, and lexical [1]. There are a variety of methods available to compute word similarity or relatedness. They can be grouped into two methods:. The first involves groups or categories into which the concepts expressed by words take up a natural position. The second concerns the position in which words occur in phrases and which sequences are more likely to occur than others. According to Hirst and St-Onge, the approach goes beyond simple edge-counting and takes into account a broader context within the full vector of words and in relation to anomalies in language that can extend the number of links [3]. Methods by Random, Wu–Palmer, and Leacock–Chodorow tests of similarity, return character strings, relative depth or paths [10,6], and density [2]. Interest in statistical and machine learning approaches, as opposed to analytical methods, is increasing and it was suggested by Resnik, Lin, and Jiang–Conrath to combine knowledge sources, such as a thesaurus, with basic corpus statistics [9, 8, 5]. The existing approaches to Word sense disambiguation (WSD) are categorized according to the primary source of knowledge employed within the procedure for the differentiation of sense. The methods that have been acknowledged in dictionaries, thesauri and lexical knowledge bases that do not incorporate any form of corpus proof are defined as dictionary-based or knowledge-based methods. With regard to the type of evidence or knowledge sources utilized, the existing algorithms relating to monolingual WSD are clustered within two major groupings, namely knowledge-based approaches and machine learning-based approaches, of which the former are further categorized into supervised, unsupervised and semi-supervised approaches [13]. The supervised approaches utilize a sense tagged corpus, the unsupervised approaches utilize an untagged corpus and the semi-supervised approaches utilize a limited amount of tagged corpora but incorporate large quantities of untagged corpora. The first algorithm that was developed in relation to semantic disambiguation was the Lesk algorithm (1986). It was applied to all words and there was no restriction or preparation phase carried out on the text before the algorithm was applied. The concept behind this algorithm was to identify where different senses overlapped and thereby to understand which words were most associated with disambiguation. This was carried out by first identifying the number of words which each sense had in common. The pairs of words from each sense which had the highest number of overlapping occurrences were then selected. A sense was then assigned to each word pair. Ambiguous word pairs were manually interpreted by definitions from the Oxford Advanced Learner's Dictionary. It was observed that this algorithm was able to identify with 50–70% precision the different senses, indicated by the word pairs [7].

Many other concepts of relatedness in WordNet introduced as apart from the is–a relation [5]. It was intended to assess the connectivity between heterogeneous pairs of parts of speech, for example, the relatedness between a noun and a verb. On this the strength of all semantic relatedness measurements would rely. It was originally used to identify lexical chains, which are a series of related words that maintain coherence in a written text. The algorithm was evaluated using the Senseval-2 English lexical sample data. Each of the 4,328 instances consists of a sentence with one target word to be disambiguated. Additional context comes from one or two surrounding sentences

An adapted Lesk algorithm was proposed [3]. The probably sense in a particular context is identified from definitions of target and related words. The combination of senses in a text is scored using a function, to identify the sense configuration with the highest score. The adapted Lesk algorithm uses the WordNet hierarchy to expand the context of a target word by considering hypernyms, hyponyms, holonyms, meronyms, troponyms, attribute relations, and their associated definitions. When a comparison was made on 4,320 ambiguous instances in the Senseval-2 English noun data set, the precision of the algorithm doubled to 32%.

In order to calculate the similarity between senses in WorldNet, researchers proposed combining domain information and the Wu-Palmer similarity measure. The genetic word sense disambiguation algorithm (GWSD) was first tested on two sets of domain terms. Almost all the terms were successfully disambiguated. The next step was to develop a new fitness function that disambiguates terms by weighting the frequency of usage using the weighted genetic word sense disambiguation algorithm (WGWSD). It was tested on SemCor which was extracted from the Brown Corpus and tagged semantically with WordNet senses. Based on nouns from 74 SemCor files, using the GWSD algorithm some researchers achieved 64.2% precision. However, when the WGWSD algorithm was used on the same set, the best and worst precision reported were 83.51% and 56.83% respectively, depend on the files used. The average precision recorded by researchers was 71.98% [11].

2. Research Method

There are three traditional semantic relatedness sequential approaches applied in computational linguistics, which can be used to measure and resolve problems associated with word sense disambiguation. These are summarised in figure 1.

2.1 Pre-Processing Phase

The first is the Pre-Processing Phase, which is the most important process as it prepares a summary of the text and analyses the structure. The level of efficiency at this stage will affect accuracy in the later stages. This phase can be broken down into three sub-processes, which are Tokenisation, Stop Word Removal, and Stemming.

- **Tokenization**:- it is based on white space and punctuation and divides the text into sentences and words.
- **Stop words removal**:- in this step it examines the text from the perspective of words that are redundant in the computational analysis. This includes prepositions (on, at, over), questions (if, do, how), and auxiliary verbs (can, could, might).... Etc.
- **Stemming**:- the third step is Stemming, which applies the Porter approach by removing suffixes and prefixes and reducing a word to its canonical form. The algorithm used distinguishes consonants and vowels in this process.

The following example show pre-processing steps on Quranic verse (Ayah):

Quranic Ayah: *2. Praise be to Allah, the Cherisher and Sustainer of the worlds.*

Tokenizaion result:

“2””.“”Praise“”be“”to“”Allah“”,“”the“”Cherisher“”and“”Sustainer“”of“”the“”worlds“”.“

Stop Words Removal Result: *“Praise””Allah””Cherisher””Sustainer””worlds”*

Stemming Result*:*

“Praise””Allah””Cherish””Sustain”

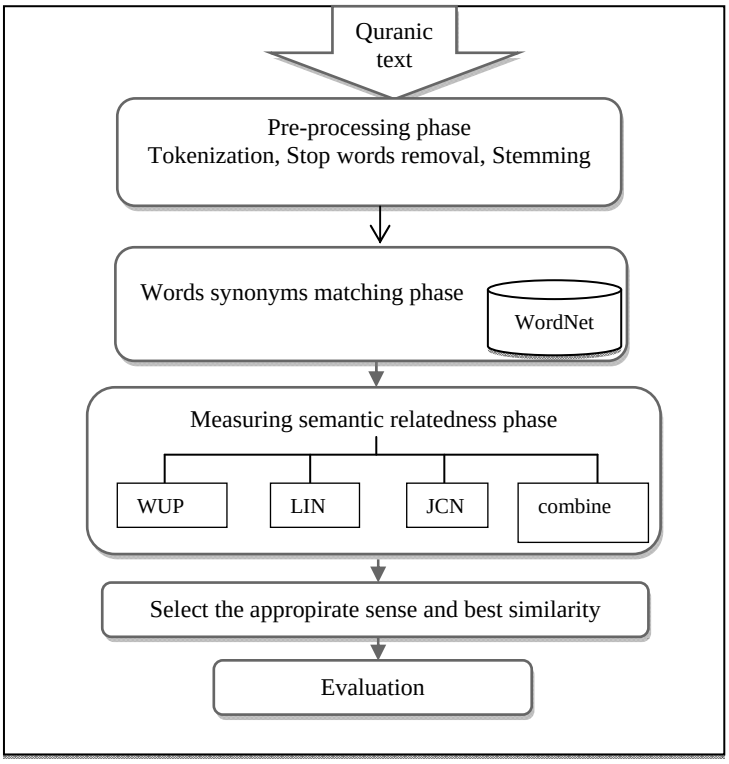

Figure 1. Framework of the Quranic WSD

2.2 Words Synonyms Matching Phase

In this phase, Synonyms and Word Matching are processes that involved. The WordNet dictionary is used to establish all the possible meanings and to select the best similarity to words used in the Quranic text. The algorithm is able to identify all the potential senses that could involve the target word and the words that appear immediately before and after the target word are from the window of context.

Table 1. Provides an example of the WordNet synonyms for Surat-al-Fatihah

Word	Synonyms
Allah	God, Lord, Sustainer, Master
Gracious	Beneficent, Affectionate
Merciful	Dispenser of Grace
Judgement	Recompense, Requital, religion
Way	Road, Path, Route

2.3 Semantic Similarity Measuring Phase

In this phase, it which examines word strings or syntax, to score the possible meanings. This phase is divided into identifying the relative depth of semantic similarity, and the information content based methods. The scoring assesses the different meanings and senses of a target word and relates it to the senses in the surrounding words. The depth relative method used in our study considers a target word and the shortest path length between two sense nodes or semantic distance. To quantify similarity, it also considers the depth of the edges and connectivity to the structure of the ontology. An example of this is the Wu–Palmer Similarity Measure. The Informational Content approach quantifies the amount of information that is associated with each sense, and the values intermediate senses in the taxonomy range from 1 to 0. A leaf node word will score 1, as it cannot be further associated or disassociated within its

context. However, at the root node level the sense can be have more than one linkage and has the most abstract level of meaning and scores 0. The Lin- and Jiang–Conrath Similarity Measures are example of this approach. The methods we have adopted are as follows:

1) **WU Palmer (WUP):** the Wu–Palmer test of senses (S1 and S2) to determine features shared by the two sense nodes, considering the depths of sense nodes in the ontology [10] and the longest common subsumer (LCS).

$$sim(s1, s2) = \frac{2 \times Depth\ ((LCS(s1, s2))}{Depth(s1) + Depth\ (s2)} \quad (1)$$

2) **Lin (LIN):** Lin tests, based on the similarly of the informational content (IC) which is found in the specific ancestor node and measures the closeness in concept [8].

$$sim_{Lin}(S_1, S_2) = \frac{2 \times \log P\big(\mathrm{LCS}(S_1, S_2)\big)}{\log P(S_1) + \log P\ (S_2)} \quad (2)$$

3) **Jiang Conrath (JCN):** the Jiang–Conrath similarity test, which examines the juxta positioning of semantic and informational content (IC), which can assess each edge to find the maximum similarity and use statistical probability to overcome the unreliability of edge distances [5].

$$\begin{aligned} Distance_{JCN}(S_1, S_2) &= IC(S_1) + IC(S_2) - 2 \times IC\big(\mathrm{LCS}(S_1, S_2)\big) \\ &= 2 \log P\big(lso(S_1, S_2)\big) - (\log P(S_1) + \log P(S_2)) \\ sim\ JCN(s1, s2) &= \frac{1}{Distance} \end{aligned} \quad (3)$$

4) **Combination Method:** In the combination of similarity measures it takes the advantages from the methods mentioned above in order to improve the quality of similarity result. The similarity between two senses $s1$ and $s2$ will be computed according to the previous eqequations, thenilter the result by applying a very common statistical process which is smoothing factor to give a weight for each method by applying the following Equation:

$$\mathrm{sim}_{\mathrm{overall}}(\mathrm{S}_1, \mathrm{S}_2) = \lambda_1 \times \mathrm{sim}_{\mathrm{WP}} + \lambda_2 \times \mathrm{sim}_{\mathrm{Lin}} + \lambda_3 \times \mathrm{sim}_{\mathrm{jcn}} \quad (4)$$

Where $0 < \lambda < 1$. Moreover, $\lambda_1 + \lambda_2 + \lambda_3 = 1$.
The weight λ is given to each method to improve the quality of similarity result. Where Simwup(s1,s2) result of the wu palmer measure, $\mathrm{sim}_{\mathrm{L}}(\mathrm{S}_1, \mathrm{S}_2)$result of Lin similarity measure and $\mathrm{dist}_{\mathrm{JCN}}(\mathrm{S}_1, \mathrm{S}_2)$ result of Jiang and Conrath measure.

3. Evaluation and Experimental Results

The purpose of word sense disambiguation in the context of this study is based on target words that appear in data prepared from the Quranic texts. The first step is to retrieve from WordNet those words which have ambiguous senses. Next, an algorithm is applied to the selected word window which takes into account all the words preceding and all subsequent words to the target word found in WordNet. Once the juxta positioned words are identified, they are also assessed in relation to the potential sense they convey. The relatedness is measured by comparison of the surrounding words to the semantic context of the target word. Finally, a computation is made of the scores for the sense of the target word against the senses of the surrounding words. The highest score is selected as it will indicate which is the most likely candidate sense based on its relevance to the context. The empirical evaluation of the Quran is based on the Budanitsky and Hirst model [1]. The approach considers spelling sensitivity in relation to nearby words and the semantic relatedness of the different spellings. This follows on from the Wu–Palmer, Lin, and Jiang–Conrath tests of similarity already described, and it takes into account whether spelling anomalies are clearly related to existing semantic concepts. The measurement in this system is based on a comparison of word pairs. The number of instances the relationship is presumed to be accurate, is divided by the total number of instances. The

verses are taken from the English translation and the results are shown in Table 2. The specific relatedness based on each approach is defined in separate columns in the Table. From the scores, the Combination of approaches (novel method) is ranked the highest because it employes the advantages of all measurements, followed by the Wu Palmer, Lin, and Jiang–Conrath approaches. And figure 2. Represents the correlation results between the proposed methods result and combined method results.

Table 2. The comparison on the similarities measures on accuracy performance

Ayah citation	Target Word	Window of context	Best suitable meaning	WUP	LIN	JCN	Combine them
(2:268)	poverty	Evil, threatens, poverty, bids, conduct, unseemly, promiseth, forgiveness, bounties, careth, knoweth.	Poverty#1: the state of the state of lack of money and material possessions	75.0	71.4	66.6	88.8
(108:2)	prayer	Lord, prayer, sacrifice.	Prayer#1: (significant or urgent request) "an entreaty to stop the fighting"; "an appeal for help"; "an appeal to the public to keep calm"	47.8	45.8	44.0	83.3
(2:285)	messenger	Messenger, believeth, hath, revealed, lord, believed, angels, books, messengers, distinction, messengers, obey, seek, forgiveness,lord, journeys.	Messenger#1: a person who carries a message	66.7	66.6	65.6	70.0
(23:16)	judgment	Day, judgment, raised.	Judgment#7: The mental ability to understand and discriminate between relations)	76.9	77.5	66.3	80.1

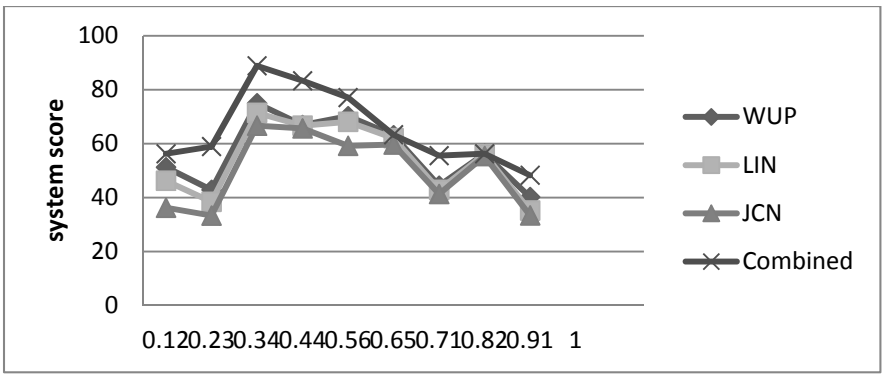

Figure 2. The correlation results between the combined method result and semantic relatedness measurements

4. Conclusion

The application of WSD in IR is has been risen due to the availability of Word Sense information like WordNet. Our article on Word Sense Disambiguation uses three traditional of semantic relatedness measurements and applied on English Quranic Translation as a DataSet which based on Abdullah Yusuf Ali (YA) [12]. His translations cover a large number of readers of the Quran in the English language. This article motivation to solve the problem of ambiguity words in English Al Quran in order to help people who are not familiar with the Arabic language to understand Allah's guidance. This study achieves a better result in the novel method which is a combination of these measurements and we can also make as a future work i) a combination

of these measurements with supervised or unsupervised word sense disambiguation methods, or ii) make a combination with structural semantic interconnection (SSI) because SSI used to create a structural specifications to the expected significances of the candidate senses for each word in a context. SSI can also be used for different semantic disambiguation problems such as disambiguate sentence in general texts, disambiguate words in glossary definitions and automatic ontology population

Acknowledgement

This research project is funded by Malaysian Govrnment under research grant ERGS/1/2013/ICT07/UKM/03/1.

References

[1] A Budanitsky and G Hirst. *Semantic distance in WordNet: An Experimental, application-oriented evaluation of five measures*. In Workshop on WordNet and other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics. Pittsburgh. 2001.

[2] Eneko Agirre and German Rigau. *Word sense disambiguation using conceptual density*. In Proceedings of the 16th International Conference on Computational Linguistics. Copenhagen. 1996: 16–22.

[3] Banerjee S and T Pedersen. *An adapted Lesk algorithm for word sense disambiguation using WordNet*. Computational linguistics and intelligent text processing, Springer. 2002: 136-145.

[4] Hirst G and St-Onge D. *Lexical Chains as representations of context for the detection and correction of malapropism*. In Fellbaum 1998: 305-332.

[5] Jiang J and Conrath D. *Semantic similarity based on corpus statistics and lexical taxonomy*. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan. 1997.

[6] Leacock C and Chodorow M. *Combining local context and WordNet similarity for word sense identification*. In Fellbaum. 1998: 265-283.

[7] Lesk and Michael. *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*. in Proceedings of the 5th annual international conference on Systems documentation, ACM. 1986.

[8] Lin D. *An information-theoretic definition of similarity*. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI. 1998.

[9] Resnik P. *Using information content to evaluate semantic similarity*. In Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal. 1995: 448-453.

[10] Wu Z and Palmer M. *Verb Semantics and Lexical Selection*. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico. 1994.

[11] Zhang C. et al., *Genetic word sense disambiguation algorithm*. In Intelligent Information Technology Application, IITA'08, Second International Symposium on, IEEE. 2008.

[12] Abdullah Yusuf Ali (YA). The meaning of the Holy Qur'an Text. Amana Publication, New Edition Translation, 10th Edition. First published in 1934, reprinted in 2003.

[13] Brown SW Dligach D & Palmer M. *VerbNet class assignment as a WSD task"*. In Computing Meaning. Springer Netherlands. 2014: 203-216.

[14] Abualkishik A & Omar K. Quran vibrations in Braille code. *Electrical Engineering and Informatics*. *ICEEI'09. International Conference on. IEEE*. 2009: 1.

[15] Abualkishik A & Omar K. *Framework for translating the Holy Quran and its reciting rules to Braille code*. In Research and Innovation in Information Systems (ICRIIS), 2013 *International Conference on. IEEE*. 2013: 380-385.

[16] Al-Bulushi SKH. *The Translation of the Names of Allah Mentioned in the Al-Qur'an Into English*. Ph.D dissertation, Universiti Sains Malaysia. 2009.