Word Sense Disambiguation Based on Semantic Relatedness Measurements

Sarah Abdul-Ameer, Sabrina Tiun, and Nazlia Omar

Abstract—The methods and background introduced in this report, concern building an understanding of the interpretation of an English translation of the Quranic text, using word sense disambiguation. Basing the research on WordNet and various similarity measures, it examines potentially misaligned or ambiguous words. Three measures of similarity are applied to the identified senses, to carry out a comparison and detailed evaluation. It is found that the Wu–Palmer approach to determine the similarity of senses is the most accurate, followed by the Lin, and Jiang–Conrath similarity measures.

Index Terms—Quranic Translated Text; Semantic Relatedness; Quranic IR; WordNet.

1 Introduction

THE central religious text of Islam is the Al-Quran, a holy book that conveys teachings and guidance on conduct and rules that should be followed. 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 are 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 verses 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 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. However, a review of the literature found no reports on WSD used in the context of the Quranic IR. 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 [1,4].

There are a variety of methods available to compute word similarity or relatedness. They can be grouped into t two methods used to compute the inter-relatedness of words. 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 [6]. Methods by Random, Wu–Palmer, and Leacock–Chodorow tests of similarity, return character strings, relative depth or paths [8,12], and density [3]. Interest in statistical and machine learning approaches, as opposed to analytical methods, is

- *Sarah Abdul-Ameer is with Faculty of Technology and Computer Science, Universiti Kebangsaan Malaysia, Malaysia, Bangi. E-mail: rosa_sara90@yahoo.com.*
- *Sabrina Tiun is with the Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Malaysia, Bangi. E-mail: sabrinatiun@gmail.com.*
- *Nazlia Omar is with Center for Artificial Intelligence Technology, University of Kebangsaan Malaysia,, Malaysia, Bangi. E-mail:- nazlia@ukm.edu.my.*

increasing and it was suggested by Resnik, Lin, and Jiang–Conrath to combine knowledge sources, such as a thesaurus, with basic corpus statistics [11,10,7].

The paper concludes with a review in Section 2 of other research and the historical development of WSD approaches.For the purposes of this research, the approach adopted is documented in Section 3, which describes the structure on which it has been based, the materials which have been taken into consideration, and the methods applied. After these steps have been presented, an evaluation of the experimental results is discussed in Section 4. The application of word sense disambiguation is expanded in Section 5, and conclusions from the foregoing discussion are interpreted in Section 6, noting the need for complementary research.

2 RELATED WORKS

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 [9].

Hirst et al. (1998) introduced many other concepts of relatedness in WordNet apart from the is–a relation. 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 instancesconsists of a sentence with one target word to be disambiguated. Additional context comes from one or two surrounding sentences [6].

An adapted Lesk algorithm was proposed by Banerjee and Pedersen et al. (2002). 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% [5].

To calculate the similarity between senses in WorldNet, Zhang and Zhou et al. (2008) 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, dependent on the files used. The average precision recorded by researchers was 71.98% [13].

3 MATERIALS AND METHODS

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.

3.1 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.

1) Tokenization: it is based on white space and punctuation and divides the text into sentences and words.

2) 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.

3) 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.

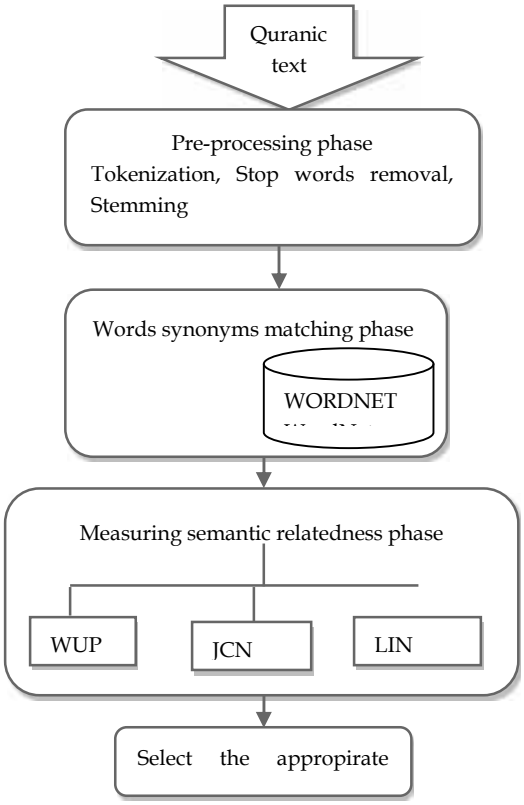


Fig. 1. System architecture

3.2 Words Synonyms Matching Phase

In this phase, which takes into consideration Synonyms and Word Matching. 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-Fatiha.

3.3 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

TABLE 1. WORDNET SYNONYMS FOR SURAT-L-FATIHA

word	Words from Surat-l-Fatiha and its synonyms *Synonyms and definitions*
Allah	Allah Muslim name for the one and only God
gracious	gracious disposed to bestow favors; "thanks to the gracious gods"
Gracious	courteous, gracious, nice exhibiting courtesy and politeness; "a nice gesture"
Gracious	benignant, gracious characterized by kindness and warm courtesy especially of a king to his subjects; "our benignant king
Gracious	gracious Characterized by charm, good taste, and generosity of spirit; "gracious even to unexpected visitors"; "gracious living"; "he bears insult with gracious good humor
Merciful	merciful used conventionally of royalty and high nobility; "gracious; "our merciful king
Praise	praise offering words of homage as an act of worship; ""they sang a hymn of praise to God
Praise	praise, congratulations, kudos, extolment an expression of approval and commendation; "he "always appreciated praise for his work
Sustainer	upholder, maintainer, sustainer someone who upholds or maintains; "firm upholders of tradition"; "they are sustainers of the "idea of democracy

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:

- 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 [12] and the longest common subsumer (LCS).

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

- 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 [10].

$$simLin = \frac{2\times logP(LCS(s1,s2))}{logP(s1)+logP(s2)} \quad (2)$$

- Jiang Conrath (JCN): the Jiang–Conrath similarity test, which examines the juxta

positioning of semantic and informational content (IC), which can assesses each edge to find the maximum similarity and use statistical probability to overcome the unreliability of edge distances [7].

$$distanceJCN(s1,s2) = IC(s1) + IC(s2) - 2 \times IC))$$
$$= 2LogP\{LSO(s1,s2)\} - \{LogP(s1) + LogP(s2))$$
$$simJCN(s1,s2) = \frac{1}{Distance} \quad (3)$$

4 EVALUATION AND EXPERIMENTALRESULTS

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.

5 APPLYING WORD SENSE DISAMBIGUATION

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 the three words preceding and three subsequent words to the target word found in WordNet. Once the juxtapositioned 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.

6 CONCLUSION AND FUTURE WORK

The three traditional semantic relatedness tests using the Wu–Palmer, Lin, and Jiang–Conrath measurements have been used to examine different Ayah (verses) from the holy Quran translated by most popular translation as a dataset based on Abdullah Yusuf Ali (YA) [2]. His translations cover a large number of readers of the Quran in the English language. The verses are taken from the English translation and the results are shown in Table 2. The specific relatedness based on each approach is

TABLE 2. ACCURACY OF WUP, LIN, JCN

Target word	Semantic relatedness measurements			
	verses	*wup*	*lin*	*jcn*
Poverty	Surah 2, verse (286)	75.00	71.429	66.66
Prayer	Surah 108,verse (2)	47.82	45.83	44.00
Messenger	Surah 2, verse (285)	66.7	66.676	65.66
Judgements	Surah 23,verse(16)	76.92	77.5	66.3

ACKNOWLEDGMENT

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, June 2001.*

[2] 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.

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

[4] Eggebraaten, T. J., et al. (2014). Natural language processing ('NLP'), Google Patents.

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

[6] Hirst G., St-Onge D., Lexical Chains as representations of context for the detection and correction of malapropisms, In Fellbaum 1998, pp. 305-332.

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

[8] Leacock C., Chodorow M., Combining local context and WordNet similarity for word sense identification, In Fellbaum 1998, pp 265-283.

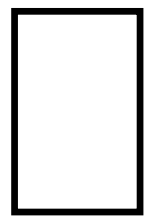
[9] Lesk, Michael. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. " *Proceedings of the 5th annual international conference on Systems documentation.* ACM, 1986.

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

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

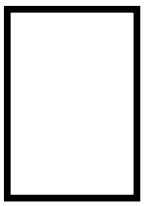
[12] Wu Z., Palmer M, Verb Semantics and Lexical Selection, *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico, 1994.

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

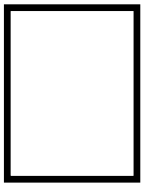


Sarah Abdul-Ameer obtained her bachelor degree in computer science from university of Baghdad, Iraq, Currently; she is a Masters student at the School of Information Technology, Faculty of computer Science and Technology, Universiti Kebangsaan Malaysia. Her research interests include Natural Language Processing and

Computational linguistics.



Sabrina Tiun obtained her PhD in Natural Language Processing (Speech Processing) and Master of Computer Sciences from Universiti Sains Malaysia, Penang, Malaysia. She is currently a Senior Lecturer a senior lecturer from Universiti Kebangsaan Malaysia Her research interests range from Speech Processing, Natural Language Processing and Information Retrieval.



Nazlia Omar obtained his bachelor degree in computer science from university of Manchester Institue of science and technology, England and master from university of Liverpool and PhD from Ulster university. She is Assoc. Prof at center of Artificial Intelligence and Technology at universitiKebangsaan Malaysia, Malaysia.