Experimental Study of Semantic Similarity Measures on Arabic WordNet

Nababteh Mohammed[1], **Deri Mohammed**[1]

Computer Center, FESA University, Jordan

Abstract

There are several semantic similarity measures that have been used to measure and quantify how much two concepts are alike. However, these measures have been tested, verified and compared in English language, using WordNet (WN). Few concerns have been given to study the impacts of traditional semantic similarity measures on Arabic language, embodied in Arabic WordNet (AWN). This paper aims at investigating the ability of applying semantic similarity measures over AWN and their applicability on Arabic-related applications. Having semantic measures for Arabic language will support many Arabic-based natural language processing applications. In this paper the experimental study was applied on seven semantic similarity measures from numerous semantic similarity measures. The experiments show that Wup measure has achieved the highest correlation with human ratings and the lowest value of MSE. This indicates that the Wup measure has the best performance in calculating the similarity of Arabic word pairs using AWN ontology against the other measures. In the other hand, path measure has the worst performance, because of the lowest correlation with human ratings and the highest value of MSE that it has achieved.

Keywords:

Semantic similarity, semantic similarity measure, WordNet, Arabic WordNet (AWN), AWSS.

1. Introduction

Rapid growth of developing traditional Arabic natural language processing (ANLP) and Arabic information retrieval applications created the needs to explore well defined semantic similarity measures over Arabic representational vocabulary known as Arabic ontology [1][2][3]. Semantics similarity is acquired by mapping an input text, as words and short texts into an ontology at which these words are getting their semantics by their relation represented in that ontology. To enable the discovery of such relation, several semantic similarity measures have been proposed in the literature.

The semantic measures have been proposed to compute the similarity between a pair of concepts in the structured model of the ontology [4]. Then, these measures have been used to discover the similarity between words in a free text in order to support natural language processing (NLP) and information retrieval (IR) applications. Many researchers have studied semantic similarity measures over English ontologies. However, there is lack of researches that focus on Arabic ontology. The interest of the improvement of how to find relevant information in a language other than English is growing, specifically on the collections of information written in Arabic [5]. Developing new semantic similarity measures over Arabic ontology will improve finding relevant information in Arabic language

Arabic is a very rich and complex language. Handling Arabic language in NLP and IR field is hard task. It is difficult to apply the same English language processing techniques on Arabic language. Arabic letters are written from right to left [6]. These letters take different forms based on their location in the word. Diacritics are written above or below the letters to represent the desired sound and to give a word the desired meaning. Also Arabic words show a complex internal structure, where words often incorporate affixes that mark grammatical inflections and diacritics to express different parts of speech [7].

The reminder of the paper is organized as follow: section 2 introduces the WordNet and Arabic WordNet. Semantic similarity measures selection are presented in section 3. Section 4 describes the process of selection the arabic dataset benchmark. Section 5 provides experimental study of applying the selected measures on AWN. A conclusion is presented in section 6.

2. WordNet and Arabic WordNet

WordNet is the product of a research project at Princeton University [8]. According to Meng, Huang, & Gu [9] WordNet is a large lexical database of English. It is a model for describing the concepts and relationships between them in a hierarchical way. Nouns, verbs, adverbs and adjectives in WordNet are organized by set of semantic relations into synonym sets (synsets), which represent one concept. Examples of semantic relations used by WordNet are synonymy, autonomy, hyponymy, member, similar, domain and cause and so on. Some relations are used for concept form relation and others for semantic relation. These relations represented as a hierarchy structure, which makes it a useful tool for computational linguistics and natural language processing

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

[10]. WordNet is used by many researchers to measure the semantic similarity or relatedness between a pair of concepts, since it organizes nouns and verbs into hierarchy way.

Black, Elkateb, Rodriguez, and Alkhalifa [11] developed Arabic WordNet (AWN) which is a lexical resource for modern standard Arabic (MSA) following the development process of Princeton WordNet for English. AWN enables translation on the lexical level to English and dozens of other languages [12]. AWN 2.0 was released in January of 2008, it contains 9,698 concepts, corresponding to 21,813 MSA words, and 6 different relation types, totaling 143,715 links. A later version of AWN, 2.0.1, was also released and contained 11,269 synsets, corresponding to 23,841 words, and 22 link types, totaling 161,705 links. AWN synsets belong to one of 5 parts of speech: noun (6,438), verb (2,536), adjective (456), adjective satellite (158), and adverb (110) [13]. AWN used in many ANLP and Arabic information retrieval applications to find common characteristics between concepts. This research will be based on AWN to implement the semantic measures and calculate similarity score between concepts.

3. Semantic Similarity Measures Selection

There are many semantic similarity measures based on WN to compute the semantic similarity between two concepts. These measures are divided into four categories, the path-based measures, information content measures, feature-based measures and hybrid measures [4]. In this research seven well-known measures from three categories (path-based measures, information content measures and hybrid measure) are selected to study their applicability on AWN. The feature-based measures use the glosses of the concepts which are provided in WN [9]. However, these glosses are not available in AWN, therefore feature-based measures will not be applied in this research. The selected measures in this paper are:

1. Wup: is path based measure uses the distance between concepts and the depth of the LCS in the taxonomy to compute the semantic similarity.[14]
2. Path measure: is path-based measure uses the length of the path between concepts to computer the semantic similarity [15].
3. LCH: is path-based measure uses the length of the path between concepts and the max depth of the taxonomy [16].
4. LI: is path-based measure uses non-linear equation function based on the length between concepts and the depth of the concepts in the taxonomy [17].
5. AWSS: is Arabic path-based measure uses LI formula to compute semantic similarity with modification on the depth and length computation to be proper for AWN [18].
6. $Res_{Meng.:}$ is node-based measure, also known as information content measure. In this measure we compute the IC using corpus independent method called IC_{meng} [19][20].
7. Zhou: is hybrid measure, uses two different measures families, path based measures and information content measures.[21]

The above seven measures consist of three path-based measures, two non linear path-based measures and one information content measure and one hybrid measure. The first three measures are linear path-based measures, and they are selected because they achieve good performance against other measures. The fourth measure LI is selected because it is non-linear path based measure, as well as it is the reference measure of AWSS. Fifth measure AWSS is selected because it is the first Arabic semantic similarity measure, and to compare its result on Arabic dataset against the results of the other six measures. AWSS proposed by Almarsoomi, et al. calculated the similarity between concepts using information sources extracted from AWN, which are length and depth. They used a previously developed Arabic word benchmark dataset [7] to evaluate AWSS measure by calculating word similarity on an Arabic word set with human judgments. The authors state that the experimental evaluation indicates that the Arabic measure is performing well. It has achieved a correlation value of 0.894 compared with the average value of human participants of 0.893 on evaluation dataset [18]. As shown previously the sixth measure is corpus independent measure, there are various corpus dependent measures, but we didn't use them due to the ambiguous and sparse data problem. Seventh measure is selected because it represents hybrid measure category.

4. Arabic Dataset Benchmark Selection

In this research Arabic dataset benchmark used is called AWSS benchmark. This dataset was created by Fazza et al (2012). The Arabic dataset uses the same procedures which were followed in creating English dataset benchmarks for semantic similarity. The most two common benchmark datasets are Rubenstein & Goodenough R&G [22] and Miller & Charles (M&C). To the best of our knowledge there are no Arabic benchmark datasets for semantic similarity except AWSS by Fazza et al [7].

The AWSS benchmark dataset was prepared mainly in two steps, first, determine the Arabic word pairs set, second, specify human similarity rate for word pairs. The AWSS benchmark creators fundamentally used the dataset of Rubenstein & Goodenough R&G [22]. Fazza et al created a list of Arabic word pairs contains 70 item [7].

They follow the same steps of R&G, 27 Arabic categories were created and employed to select the stimulus Arabic word pairs and to promote the best possible semantic representation. Arabic categories were created based on Rubenstein & Goodenough method, the list of English words in the R&G experiment contains 48 nouns from 22 different categories. In AWSS benchmark another five categories added to expand 22 categories to be 27 categories. The 48 English noun pairs from R&G list have been used to create the 22 Arabic categories after translated into Arabic language using English-Arabic dictionary and checked their accuracy from professional translator and fluent lecturers, the categories specified based on the definition of the selected pairs [22]. After the 22 categories were specified, new 5 categories were added, the added categories relevant to Arabic life style. After that, the first two nouns from each category are selected to generate 56 stimulus Arabic words [7].

The 56 noun pairs were divided into two columns, 28 nouns in each column. A sample of 22 Arabic native speakers from 5 different Arabic countries was chosen to generate two sets of Arabic noun pairs ranging from high similarity of meaning (HSM) to medium similarity of meaning (MSM) and low similarity. The participant asked to write 28 Arabic noun pairs which have high similarity from the list by selecting one noun from Column A and other from Column B, and write 32 pairs have medium similarity by the same procedure of selecting high similarity pairs. The participants while selecting can choose the same word more than one time without duplicating the pairs. After the list processed the final list was contains 57 Arabic noun pairs. Then 13 Arabic noun pairs from low similarity were randomly selected by Fazza et al. in order to get list from 70 Arabic word pairs which covered high to low similarity, this list called AWSS benchmark. Table 1 shows AWSS list.

Another 60 participants from different Arabic countries who had not taken part in generating Arabic word pairs were asked to rank the set of 70 Arabic word pairs previously collected. The participants were requested to rate each word pair based on how similar they were in meaning from 0.0 to 4.0 [7]. In this work, the human rating is divided by four to convert the rating from [0-4] range to [0-1]. In this paper AWSS benchmark dataset has been chosen for various reasons as follows: first, Arabic word pairs were created carefully. Second, this benchmark was based on R&G dataset, which is the most influential word dataset for English. The original Arabic dataset contained 24 low similarity, 24 medium similarity and 22 high similarity word pairs. Due to absence of some words in AWN, only 40 word pairs were taken. Sub dataset in this experiment contains 12 word pairs low similarity, 13 word pairs medium similarity and 15 high similarity word pairs.

5. Experimental Study of Applying the Traditional Measures on AWN

In this section we will study the possibility of using the traditional semantic similarity measures on Arabic ontology. The results of this study will give the researchers in Arabic natural language processing good knowledge about the semantic similarity measures that could use in AWN.

The experiments study in this section is organized as following steps, choosing the proper tools for applying the seven semantic similarity measures over AWN, applying the traditional semantic similarity measures using the selected tool, extracting and analyzing the results of implementing the measures, finally evaluating the results based on MSE and correlation.

5.1 Selecting the Optimal Tool

There are many available tools that implement the semantic similarity measures on WN. In this research the Java AWN API and WS4J will be used. Java AWN API contains implementations of four semantic similarity measures, Wup, LCH, LI and path. Additionally it gives information sources like number of hyponyms for concepts, depth of the concepts in the taxonomy and path length between concepts. Therefore, in this research we apply the four mentioned measures as well as additional measure called Resnik which based on the information provided from the tool. WS4J is the second tool used to compute the semantic similarity on English noun pairs. This tool can compute the similarity score using eight measures over WN, which easy to use online tool.

Table 1: AWSS dataset benchmark [7]

	Word Pairs		Human Ratings	أزواج الكلمـات			Word Pairs		Human Ratings	أزواج الكلمـات	
1	Coast	Endorsement	0.03	ساحل	تصديق	36	Slave	Lad	1.77	عبد	فتى
2	Noon	String	0.03	ظهر	خيط	37	Journey	Bus	1.83	رحلة	باص
3	Cushion	Diamond	0.06	مسند	الماس	38	Girl	Odalisque	1.96	فتاة	جارية
4	Gem	Pillow	0.07	جوهرة	مخدة	39	Feast	Fasting	1.96	عيد	صيام
5	Stove	Walk	0.07	موقد	مشي	40	Coach	Means	2.07	حافلة	وسيلة
6	Cord	Midday	0.08	حبل	ظهيرة	41	Brother	Lad	2.15	أخ	فتى
7	Signature	String	0.08	توقيع	خيط	42	Sage	Sheikh	2.26	حكيم	شيخ
8	Boy	Endorsement	0.12	صبي	تصديق	43	Girl	Sister	2.38	فتاة	أخت
9	Boy	Midday	0.16	صبي	ظهيرة	44	Hill	Mountain	2.60	تل	جبل
10	Slave	Vegetable	0.16	عبد	خضار	45	Hen	Pigeon	2.61	دجاجة	حمامة
11	Smile	Village	0.18	إبتسامة	قرية	46	Master	Sheikh	2.66	سيد	شيخ
12	Smile	Pigeon	0.20	إبتسامة	حمامة	47	Food	Vegetable	2.78	طعام	خضار
13	Wizard	Infirmary	0.22	ساحر	مشفى	48	Slave	Odalisque	2.84	عبد	جارية
14	Noon	Fasting	0.29	ظهر	صيام	49	Run	Walk	3.01	جري	مشي
15	Hill	Pigeon	0.33	تل	حمامة	50	Brother	Sister	3.08	أخ	أخت
16	Countryside	Laugh	0.34	ريف	ضحك	51	Cord	String	3.09	حبل	خيط
17	Glass	Diamond	0.36	كأس	الماس	52	Forest	Woodland	3.14	غابة	أحراش
18	Glass	Fasting	0.38	كأس	صيام	53	Sage	Thinker	3.30	حكيم	مفكر
19	Cord	Mountain	0.54	حبل	جبل	54	Gem	Diamond	3.38	جوهرة	الماس
20	Hospital	Grave	0.83	مستشفى	قبر	55	Cushion	Pillow	3.38	مسند	مخدة
21	Forest	Shore	0.86	غابة	شاطئ	56	Journey	Travel	3.39	رحلة	سفر
22	Gem	Young woman	0.87	جوهرة	شابة	57	Countryside	Village	3.41	ريف	قرية
23	Sepulcher	Sheikh	0.89	ضريح	شيخ	58	Smile	Laugh	3.48	إبتسامة	ضحك
24	Tool	Pillow	0.99	اداة	مخدة	59	Stove	Oven	3.55	موقد	فرن
25	Coast	Mountain	1.06	ساحل	جبل	60	Coast	Shore	3.56	ساحل	شاطئ
26	Run	Shore	1.13	جري	شاطئ	61	Signature	Endorsement	3.58	توقيع	تصديق
27	Hill	Woodland	1.19	تل	أحراش	62	Tool	Means	3.68	اداة	وسيلة
28	Countryside	Vegetable	1.24	ريف	خضار	63	Noon	Midday	3.70	ظهر	ظهيرة
29	Tool	Tumbler	1.32	اداة	قدح	64	Boy	Lad	3.71	صبي	فتى
30	Master	Thinker	1.36	سيد	مفكر	65	Girl	Young woman	3.74	فتاة	شابة
31	Feast	Laugh	1.36	عيد	ضحك	66	Sepulcher	Grave	3.75	ضريح	قبر
32	Hen	Oven	1.44	دجاجة	فرن	67	Wizard	Magician	3.76	ساحر	مشعوذ
33	Journey	Shore	1.47	رحلة	شاطئ	68	Coach	Bus	3.80	حافلة	باص
34	Coach	Travel	1.60	حافلة	سفر	69	Glass	Tumbler	3.82	كأس	قدح
35	Food	Oven	1.76	طعام	فرن	70	Hospital	Infirmary	3.91	مستشفى	مشفى

5.2 Computing the Semantic Similarity Using Java AWN API

In this section the semantic similarity measures will be applied using the java AWN API on 40 Arabic noun pairs which were selected from AWSS dataset, and the result for all measures will be described, analyzed and compared with human ratings. In order to run java AWN API tool, we should import the Arabic WordNet (AWN). Arabic WordNet browser is an application available on the internet containing the Arabic WordNet database. The AWN browser gives us the ability to export its database as a file. After exporting AWN file, the exported file should be passed to the java AWN API. The java AWN API tool contains a set of methods and classes to handle AWN. The first class has been used was AWN class. This class enable us to import the AWN xml file, it takes two parameters, the first parameter is the path of AWN xml file, the second parameter is "true" or "false", to tell the API to remove diacritics (harakat) from the source, "false" parameter should be passed, in our case we need diacritics, so "true" has been passed. The following code shows how to use the class.

AWN aw= new AWN("upc_db.xml",true);

As mentioned above, we applied the selected semantic similarity measures to all Arabic word pairs in the dataset, this step took a lot of time and effort, because we need to get synset-id for all word pairs, this has been done by two steps as follows:

1. We used AWN browser to get Arabic synonyms with diacritics by typing Arabic concept in Arabic word filed, then choosing proper word sense from the list appearing in Arabic word senses box as shown in figure 1. Thus the Arabic word with diacritics copied to be used in java AWN API tool.
2. Arabic word with diacritics have been passed to java AWN API method to get **synset ID** as follows:

```
List<String>
        ItemID=aw.Get_Item_Id_From_Name("شَيْخ");
```

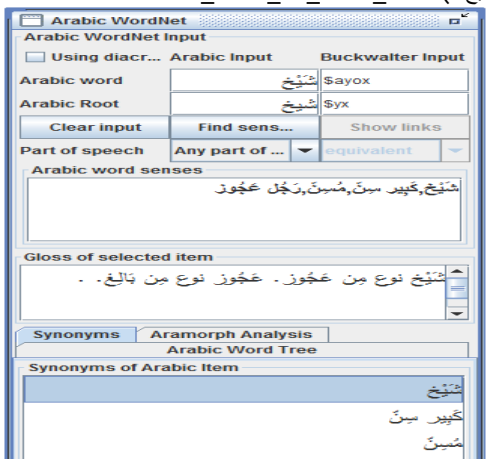


Figure 1: Arabic word senses box in AWN browser

The above two steps have been repeated for all Arabic noun pairs and all collected **synsets IDs** have have been stored.

The semantic similarity for all Arabic noun pairs have been computed by Java AWN API tools. As said previously this tool has only 4 measures, namely, **edge counting** (*Get_word_similirty_edge_counting*), **WUP** (*Get_word_similirty_wuP*), **Leakcock and Chodorow** (*Get_word_similirty_LeakcockChodorow*) And **Li** (*Get_word_similirty_Li*). For the two measures ($Resnik_{meng}$ and Zhou), we developed two new methods. Arabic word pairs were already implemented by AWSS measure (Almarsoomi et al., 2013).

To compute the semantic similarity of word pair, built-in methods in Java AWN API will be used. To do that the **synset ID** for Arabic word pairs should pass to the methods of the measures in java AWN API to return the similarity score between them. For example if we need to find the similarity score between شيخ (Sheikh) and ضريح (Sepulcher), we should pass **synset ID** for both concepts as follows:

```
System.out.println(aw.Get_word_similirty_w
uP("$ayox_n1AR","qabor_n1AR"));
System.out.println(aw.Get_word_similirty_L
i("$ayox_n1AR","qabor_n1AR",0.2,0.6));
```

5.3 Gathering the Results for All Measures and Evaluation

After calculating the similarity score for all Arabic word pairs and English word pairs using the above mentioned techniques, the next step was to gather similarity values for each measure, then study the performance for all measures. Therefore, we wrote down the results into two tables. The evaluation process in this paper was carried out by finding two factors, namely **correlation** between similarity measure score and human rating and **mean square error** (MSE) of measures results. Tables 2 & 3 show the results of applying the measures on the 40 Arabic noun pairs. Table 2 shows the results of WUP, LI and Path measures. The table contains the 40 Arabic word pairs and their translations. The Arabic word pairs have been translated into English word pairs in order to be applied over WN. The results of applying Arabic and English word pairs have been compared to study the differences between AWN and WN. The table includes **Human Rating** column which contains the human judgment similarity score of the Arabic noun pairs, this score has been used to be compared with computer based result (i.e output of applying Wup measure). Human based score is considered as benchmark to compute the error rate of the computerized semantic similarity measure. Table 2 also contains two columns (**EN**, **AR**) to show the similarity score of Wup for English and Arabic pairs. The two columns (**Err**, **Sqr_Err**) in the table contain the Error which is the difference between the computed similarity score by Wup and human rating score, and the square error to compute the mean square error. The word pairs have been divided into three groups: low similarity, medium similarity and high similarity.

Table 2 Results of applying WUP, LCH and path measures on AWN

		Word Pairs		Arabic word pairs		Human Ratings	WuP				LCH				Path			
							EN	AR	Err.	Sqr. Err.	EN	AR	Err.	Sqr. Err.	EN	AR	Err.	Sqr. Err.
1	Low Similarity	Coast	Endorsement	تصديق	ساحل	0.01	0.28	0	0.01	0.0001	0.43	0	0.01	0.0001	0.12	0	0.01	0.0001
2		Noon	String	خيط	ظهر	0.01	0.35	0	0.01	0.0001	0.33	0	0.01	0.0001	0.08	0	0.01	0.0001
3		Stove	Walk	مشي	موقد	0.01	**0.16**	-	-	-	**0.17**	-	-	-	0.04	-	-	-
4		Cord	Midday	ظهيرة	حبل	***0.02***	0.21	0	0.02	0.0004	0.25	0	0.02	0.0004	0.06	0	0.02	0.0004
5		Signature	String	خيط	توقيع	***0.02***	0.23	0	0.02	0.0004	0.28	0	0.02	0.0004	0.07	0	0.02	0.0004
6		Boy	Endorsement	تصديق	صبي	***0.03***	0.23	0	0.03	0.0009	0.33	0	0.03	0.0009	0.09	0	0.03	0.0009
7		Boy	Midday	ظهيرة	صبي	***0.04***	0.28	0	0.04	0.0016	0.33	0	0.04	0.0016	0.06	0	0.04	0.0016
8		Smile	Village	قرية	إبتسامة	***0.05***	0.37	0	0.05	0.0025	0.35	0	0.05	0.0025	0.09	0	0.05	0.0025
9		Noon	Fasting	صيام	ظهر	***0.07***	0.36	0	0.07	0.0049	0.27	0	0.07	0.0049	0.06	0	0.07	0.0049
10		Glass	Diamond	الماس	كأس	***0.09***	0.35	0.12	-0.03	0.0009	0.40	**0.22**	-0.13	0.0169	0.11	0.07	0.02	0.0004
11		Sepulcher	Sheikh	ضريح	شيخ	***0.22***	0.47	0.18	0.04	0.0016	0.35	0.35	-0.13	0.0169	0.09	0.11	0.11	0.0121
12		Countryside	Vegetable	خضار	ريف	***0.31***	0.40	0.18	0.13	0.0169	0.33	0.35	-0.04	0.0016	0.08	0.11	0.2	0.04
13	Medium similarity	Tumbler	Tool	قدح	أداة	***0.33***	0.73	0.5	-0.17	0.0289	0.52	0.43	-0.1	0.01	0,16	0.12	0.21	0.0441
14		Laugh	Feast	عيد	ضحك	***0.34***	0.40	0.15	0.19	0.0361	0.42	0.33	0.01	0.0001	0.16	0.09	0.25	0.0625
15		Girl	Odalisque	جارية	فتاة	***0.49***	0.83	0.54	-0.05	0.0025	0.57	0.59	-0.1	0.01	0.2	0.2	0.29	0.0841
16		Feast	Fasting	صيام	عيد	***0.49***	0.5	0.18	0.31	0.0961	0.33	0.22	0.27	0.0729	0.09	0.07	0.42	0.1764
17		Coach	Means	وسيلة	حافلة	***0.52***	0.77	0.66	-0.14	0.0196	0.56	0.52	0	0	0.20	0.2	0.32	0.1024
18		Sage	Sheikh	شيخ	حكيم	***0.56***	0.76	0.46	0.1	0.01	0.52	0.76	-0.2	0.04	0.16	0.14	0.42	0.1764
19		Girl	Sister	أخت	فتاة	***0.60***	0.40	0.54	0.06	0.0036	0.33	0.52	0.08	0.0064	0.08	0.2	0.4	0.16
20		Hen	Pigeon	حمامة	دجاجة	***0.65***	0.84	0.78	-0.13	0.0169	0.57	0.59	0.06	0.0036	0.2	0,2	0.45	0.2025
21		Hill	Mountain	جبل	تل	***0.65***	0.85	-	-	-	0.70	-	-	-	0.33	-		
22		Master	Sheikh	شيخ	سيد	***0.67***	0.90	0.5	0.17	0.0289	0.70	0.46	0.21	0.0441	0.33	0.16	0.51	0.2601
23		Food	Vegetable	خضار	طعام	***0.69***	0.85	0.4	0.29	0.0841	0.70	0.42	0.27	0.0729	0.33	0.16	0.53	0.2809
24		Slave	Odalisque	جارية	عبد	***0.71***	0.72	0.66	0.05	0.0025	0.47	0.68	0.03	0.0009	0.14	0.5	0.21	0.0441
25		Run	Walk	مشي	جري	***0.75***	0.90	0.83	-0.08	0.0064	0.70	0.68	0.07	0.0049	0.33	0.5	0.25	0.0625
26	High Similarity	Cord	String	خيط	حبل	***0.77***	0.94	0.66	0.11	0.0121	0.81	0.59	0.18	0.0324	0,5	0.25	0.52	0.2704
27		Forest	Woodland	أحراش	غابة	***0.79***	1	0.88	-0.09	0.0081	0.35	0.91	-0.12	0.0144	1	1	-0.21	0.0441
28		Sage	Thinker	مفكر	حكيم	***0.82***	0.85	0.8	0.02	0.0004	0.63	0.79	0.03	0.0009	0.25	0.5	0.32	0.1024
29		Journey	Travel	سفر	رحلة	***0.84***	0.95	0.90	-0.06	0.0036	0.70	0.88	-0.04	0.3598	0.5	1	-0.16	2.1363
30		Gem	Diamond	ألماس	جوهرة	***0.84***	0.95	0.83	0.01	0.0001	0.83	0.9	-0.06	0.0036	0.5	0.5	0.34	0.1156
31		Countryside	Village	قرية	ريف	***0.85***	0.77	0.80	0.05	0.0025	0.55	0.9	-0.05	0.0025	0.2	1	-0.15	0.0225
32		Cushion	Pillow	مخدة	مسند	***0.85***	0.94	0.57	0.28	0.0784	0.70	0.46	0.39	0.1521	0.5	0.16	0.69	0.4761
33		Smile	Laugh	ضحك	إبتسامة	***0.87***	0.87	0.62	0.25	0.0625	0.70	0.40	0.47	0.2209	0.33	0.16	0.71	0.5041
34		Signature	Endorsement	توقيع	تصديق	***0.89***	0.94	0.8	0.09	0.0081	0.8	0.76	0.13	0.0169	0.5	0.5	0.39	0.1521
35		Tools	Means	وسيلة	أداة	***0.92***	0.82	0.76	0.16	0.0256	0.63	0.68	0.24	0.0576	0,25	0.5	0.42	0.1764
36		Sepulcher	Grave	ضريح	قبر	***0.93***	0.94	1	-0.07	0.0049	0.80	0.68	0.25	0.0625	0.5	1	-0.07	0.0049
37		Boy	Lad	فتى	صبي	***0.93***	0.95	0.88	0.05	0.0025	0.79	1	-0.07	0.0049	0.5	1	-0.07	0.0049
38		Wizard	Magician	مشعوذ	ساحر	***0.94***	1	-	-	-	0.98	-	-	-	1	-	-	-
39		Coach	Bus	حافلة	باص	***0.95***	1	1	-0.05	0.0025	1	0.91	0.04	0.0016	1	1	-0.05	0.0025
40		Glass	Tumbler	قدح	كأس	***0.95***	0.94	0.77	0.18	0.0324	0.70	**0.59**	0.36	0.1296	0.5	0.5	0.45	0.2025
						MSE	0.016475676				0.031743243				0.160383784			

The Wup column in Table 2 shows that Wup measure has obtained a good value of MSE (0.016475). MSE values for each similarity group (i.e. low, medium and high) were calculated separately. MSE value for high similarity group is (0.01740). Low and medium similarity group have the same MSE value (0.0027). These results indicate better performance for Wup in high similarity.

Wup measure has obtained a high value of correlation coefficient (0.94) with human rating, this means that Wup measure has good linear relation with human rating. Figure 2-A shows the correlation between human ratings and the scores of Wup measure.

The LCH column in Table 2 shows that the LCH measure has obtained MSE value of (0.037075). The results show that the LCH measure performs better in low similarity group with MSE value of (0.00231). The LCH measure has the worst performance in high similarity group due to the highest value of MSE (0.06085) which this measure has achieved.

LCH measure has a good correlation coefficient compared with human rarings (0.89). This indicates a strong relation between LCH measure and human ratings. Less correlation has been scored when compared with LCH measure on WN (0.82). Figure 2-B shows the correlation between the scores of LCH measure and human ratings.

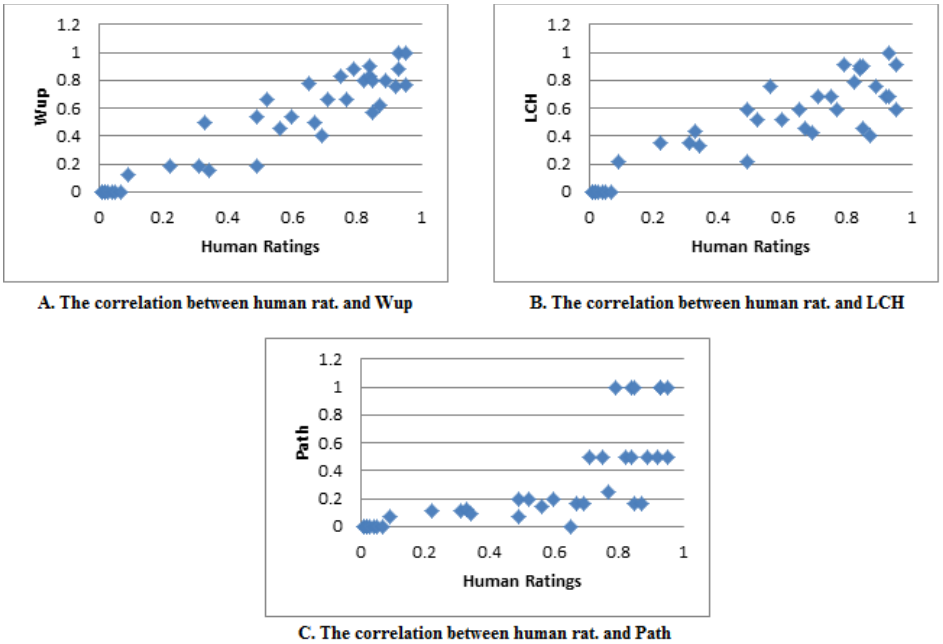
The column of Path measure in table 2 shows that Path measure has obtained the highest MSE value (0.160383) compared to the MSE values of other measures, which indicates bad performance for path measure. Highest MSE value (0.301057) for this measure in high similarity group shows that path measure has scored very poor results in high similarity.

The correlation coefficient of path measure is 0.75. Figure 2-C shows an empty area between 0.5 and 1.However, this empty area reduces the correlation with human ratings. Path measure on AWN has scored better value of correlation coefficient compared with path measure that has been applied on WN with value of (0.79).

Table 3 shows the results of the remaining four semantic similarity measures (i.e. Li, Resmeng, AWSS and Zhou). The column of Li measures in table 3 shows that MSE value for Li's measure is (0.1020513). This high value of error indicates poor performance. The results show that Li's measure has obtained better scores for low similarity group than scores for medium and high similarity group.

Correlation coefficient of Li's similarity measure using AWN beats the path measure with value of (0.84). Li's measure has scored high correlation coefficient with corresponding Li's measure that has been applied over WN with value of (0.95).

Information content-based measure (Resmeng) has obtained medium value of MSE (0.077056). Compared to the other measures. This measure has achieved intermediate performance. This measure performs well in low similarity group by achieving (0.014863) of MSE in low similarity group. However, the results show weakness of this measure in high similarity. ResMeng measure has obtained a good correlation (0.91) with human ratings and comes second place after Wup measure. Correlation value between ResMeng measure over AWN and ResMeng measure over WN is 0.82.



A. The correlation between human rat. and Wup

B. The correlation between human rat. and LCH

C. The correlation between human rat. and Path

Figure 2: The correlation between result of Wip,LCH and Path measures and human ratings

Table 3 shows that **EN** sub-column for AWSS column has no values, because this measure has been developed especially to be applied on AWN. However this measure has achieved good MSE score (0.044237). AWSS measure has scored best in low similarity group and worst results in high similarity. Human rating correlation with AWSS method (0.88) is very close to LCH correlation with human scores. Figure 3-C shows the correlation between the scores of AWSS measure and the human ratings.

The last measure that has been applied is Zhou measure, as shown in the table 3. The MSE value (0.03174) of this measure is very close to MSE of LCH measure. MSE value of (0.07202) in high similarity group indicates the weakness of this measure in high similarity group. However, Zhou measure has achieved better performance in medium and low similarity. Figure 3-D shows the correlation between Zhou measure and human ratings, this measure has a high correlation score after Wup measure (0.92).

Table 3 Results of applying Li, Res_Meng, AWSS and Zhou measure on AWN

		Word Pairs		Arabic word pairs		Huma Rating	Li				Res_Meng				AWSS				Zhou			
							EN	AR	Err.	Sqr. Err.	EN	AR	Err.	Sqr. Err.	EN	AR	Err.	Sqr. Err.	EN	AR	Err.	Sqr. Err.
1	Low Similarity	Coast	Endorsement	تصديق	ساحل	0.01	0.09	0	0.01	0.0001	0.23	0	0.01	0.0001	-	0	0.01	0.0001	-	0	0.01	0.0001
2		Noon	String	خيط	ظهر	0.01	0.09	0	0.01	0.0001	0.36	0	0.01	0.0001	-	0.17	-0.16	0.0256	-	0	-0.16	0.0256
3		Stove	Walk	مشي	موقد	0.01	0.12	-	-	-	0.23	-	_	_	-	-	-	-	-	-	-	-
4		Cord	Midday	ظهيرة	حبل	0.02	0.09	0	0.02	0.0004	0.31	0	0.02	0.0004	-	0	0.02	0.0004	-	0	0.02	0.0004
5		Signature	String	خيط	توقيع	0.02	0.16	0	0.02	0.0004	0.20	0	0.02	0.0004	-	0	0.02	0.0004	-	0	0.02	0.0004
6		Boy	Endorsement	تصديق	صبي	0.03	0.16	0	0.03	0.0009	0.23	0	0.03	0.0009	-	0	0.03	0.0009	-	0	0.03	0.0009
7		Boy	Midday	ظهيرة	صبي	0.04	0.18	0	0.04	0.0016	0.25	0	0.04	0.0016	-	0	0.04	0.0016	-	0	0.04	0.0016
8		Smile	Village	قرية	إبتسامة	0.05	0.11	0	0.05	0.0025	0.36	0	0.05	0.0025	-	0	0.05	0.0025	-	0	0.05	0.0025
9		Noon	Fasting	صيام	ظهر	0.07	0.14	0	0.07	0.0049	0.46	0	0.07	0.0049	-	0	0.07	0.0049	-	0	0.07	0.0049
10		Glass	Diamond	الماس	كأس	0.09	0.09	0.03	0.06	0.0036	0.59	0	0.09	0.0081	-	0.05	0.04	0.0016	-	0.18	-0.09	0.0081
11		Sepulcher	Sheikh	ضريح	شيخ	0.22	0.18	0.08	0.14	0.0196	0.53	0	0.22	0.0484	-	0.06	0.16	0.0256	-	0.30	-0.08	0.0064
12		Countryside	Vegetable	خضار	ريف	0.31	0.2	0.08	0.23	0.0529	0.46	0	0.31	0.0961	-	0.45	-0.14	0.0196	-	0.30	0.01	0.0001
13	Medium similarity	Tumbler	Tool	قدح	أداة	0.33	0.25	0.19	0.14	0.0196	0.64	0.25	0.08	0.0064	-	0.54	-0.21	0.0441	-	0.51	-0.18	0.0324
14		Laugh	Feast	عيد	ضحك	0.34	0.18	0.03	0.31	0.0961	0.36	0	0.34	0.1156	-	0.66	-0.32	0.1024	-	0.25	0.09	0.0081
15		Girl	Odalisque	جارية	فتاة	0.49	0.26	0.34	0.15	0.0225	0.76	0.25	0.24	0.0576	-	0.73	-0.24	0.0576	-	0.46	0.03	0.0009
16		Feast	Fasting	صيام	عيد	0.49	0.40	0.03	0.46	0.2116	0.25	0.40	0.09	0.0081	-	0.17	0.32	0.1024	-	0.40	0.09	0.0081
17		Coach	Means	وسيلة	حافلة	0.52	0.80	0.36	0.16	0.0256	0.64	0.59	-0.07	0.0049	-	0.38	0.14	0.0196	-	0.51	0.01	0.0001
18		Sage	Sheikh	شيخ	حكيم	0.56	0.66	0.65	-0.09	0.0081	0.53	0.40	0.16	0.0256	-	0.67	-0.11	0.0121	-	0.41	0.15	0.0225
19		Girl	Sister	أخت	فتاة	0.60	0.76	0.34	0.26	0.0676	0.46	0.40	0.2	0.04	-	0.37	0.23	0.0529	-	0.46	0.14	0.0196
20		Hen	Pigeon	حمامة	دجاجة	0.65	0.80	0.36	0.29	0.0841	0.76	0.81	-0.16	0.0256	-	0.89	-0.24	0.0576	-	0.46	0.19	0.0361
21		Hill	Mountain	جبل	تل	0.65	0.82.	-	_	-	0.59	-	_	_	-	-	-	-	-	-	-	-
22		Master	Sheikh	شيخ	سيد	0.67	0.76	0.28	0.39	0.1521	0.73	0.40	0.27	0.0729	-	0.67	0	0	-	0.41	0.26	0.0676
23		Food	Vegetable	خضار	طعام	0.69	0.85	0.20	0.49	0.2401	0.59	0.25	0.44	0.1936	-	0.53	0.16	0.0256	-	0.41	0.28	0.0784
24		Slave	Odalisque	جارية	عبد	0.71	0.87	0.51	0.2	0.04	0.69	0.51	0.2	0.04	-	0.93	-0.22	0.0484	-	0.58	0.13	0.0169
25		Run	Walk	مشي	جري	0.75	0.90	0.66	0.09	0.0081	0.76	0.59	0.16	0.0256	-	0.60	0.15	0.0225	-	0.67	0.08	0.0064
26	High Similarity	Cord	String	حبل	خيط	0.77	0.85	0.44	0.33	0.1089	0.69	0.51	0.26	0.0676	-	0.70	0.07	0.0049	-	0.51	0.26	0.0676
27		Forest	Woodland	أحراش	غابة	0.79	0.96	0.80	-0.01	0.0001	0.64	0.76	0.03	0.0009	-	0.82	-0.03	0.0009	-	1	-0.21	0.0441
28		Sage	Thinker	مفكر	حكيم	0.82	0.92	0.65	0.17	0.0289	0.73	0.51	0.31	0.0961	-	0.75	0.07	0.0049	-	0.76	0.06	0.0036
29		Journey	Travel	سفر	رحلة	0.84	0.96	0.96	-0.12	1.2004	0.76	0.59	0.25	0.944	-	0.87	-0.03	0.6391	-	0.79	0.05	0.4379
30		Gem	Diamond	ألماس	جوهرة	0.84	0.95	0.66	0.18	0.0324	0.81	0.59	0.25	0.0625	-	0.89	-0.05	0.0025	-	1	-0.16	0.0256
31		Countryside	Village	قرية	ريف	0.85	0.93	0.65	0.2	0.04	0.51	0.51	0.34	0.1156	-	0.82	0.03	0.0009	-	0.67	0.18	0.0324
32		Cushion	Pillow	مخدة	مسند	0.85	0.91	0.29	0.56	0.3136	0.69	0.59	0.26	0.0676	-	0.82	0.03	0.0009	-	0.79	0.06	0.0036
33		Smile	Laugh	ضحك	إبتسامة	0.87	0.95	0.24	0.63	0.3969	0.64	0.59	0.28	0.0784	-	0.29	0.58	0.3364	-	0.79	0.08	0.0064
34		Signature	Endorsement	توقيع	تصديق	0.89	0.90	0.65	0.24	0.0576	0.76	0.51	0.38	0.1444	-	0.93	-0.04	0.0016	-	0.79	0.1	0.01
35		Tools	Means	وسيلة	أداة	0.92	0.94	0.54	0.38	0.1444	0.76	0.59	0.33	0.1089	-	0.93	-0.01	0.0001	-	0.51	0.41	0.1681
36		Sepulcher	Grave	ضريح	قبر	0.93	0.96	0.69	0.24	0.0576	0.76	0.59	0.34	0.1156	-	0.82	0.11	0.0121	-	1	-0.07	0.0049
37		Boy	Lad	صبي	فتى	0.93	0.94	0.67	0.26	0.0676	0.76	0.51	0.42	0.1764	-	0.95	-0.02	0.0004	-	0.79	0.14	0.0196
38		Wizard	Magician	مشعوذ	ساحر	0.94	0.94	-	-	-	0.76	-	_	_	-	-	-	-	-	-	-	-
39		Coach	Bus	حافلة	باص	0.95	0.96	0.88	0.07	0.0049	0.76	0.76	0.19	0.0361	-	0.94	0.01	0.0001	-	1	-0.05	0.0025
40		Glass	Tumbler	قدح	كأس	0.95	0.89	0.44	0.51	0.2601	0.73	0.71	0.24	0.0576	-	0.89	0.06	0.0036	-	0.79	0.16	0.0256
						MSE	0.102051351				0.07705675				0.044237				0.031743243			

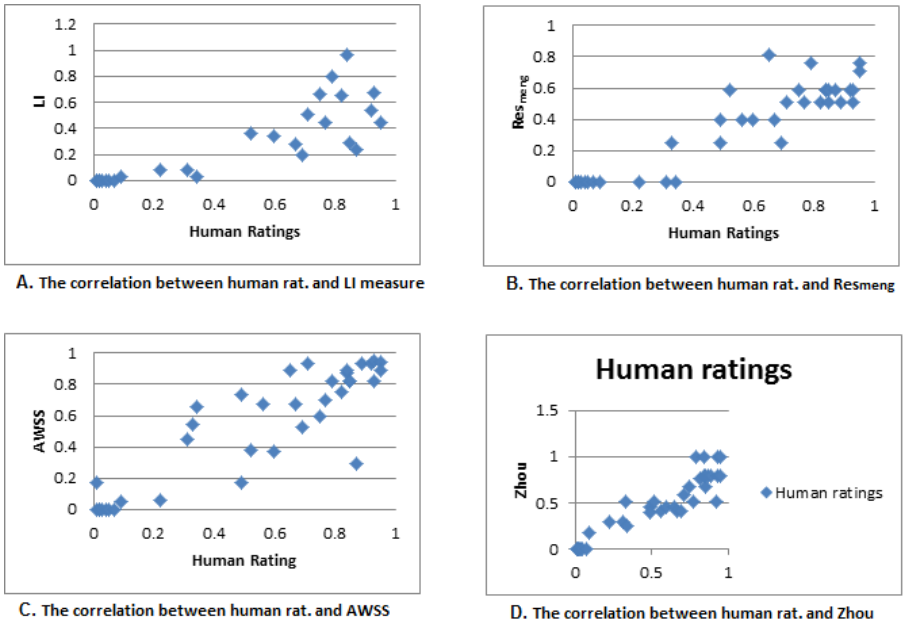


Figure 3: The correlation between human rating and LI, Resmeng, AWSS and Zhou measure

5.4 Measures Evaluation

In this section the obtained results from previous experiments have been evaluated to find which measures achieve good performance over AWN. The semantic measures performance on AWN have been compared using two factors, MSE value and correlation with human ratings.

Table 4 shows the correlation between each measure and human ratings, and the MSE values for all measures. Correlation values multiplied by 10 and MSE values multiplied by 100 to make the comparison between measures easier. Table 4 shows that Wup measure has achieved the highest correlation with human ratings and the lowest value of MSE. This indicates that the Wup measure has the best performance in calculating the similarity of Arabic word pairs using AWN ontology against the other measures. Besides, path measure has the worst performance, because of the lowest correlation with human ratings and highest value of MSE that it has achieved.

Table 4: list of correlation and MSE values for all measures

Measure	Correlation with human ratings	MSE
Wup	**9.4**	**1.6475**
Res_{Meng}	9.1	7.7056
LCH	8.9	3.7075
AWSS	8.8	4.4237
Li	8.4	10.205
Path	**7.5**	**16.038**
Zhou	9.2	3.17432

Figure 4 shows that the correlation values of all measures are almost close to each other. However, the correlation value of Wup measure is the highest, followed by Zhou measure and the correlation value of path measure is the lowest.

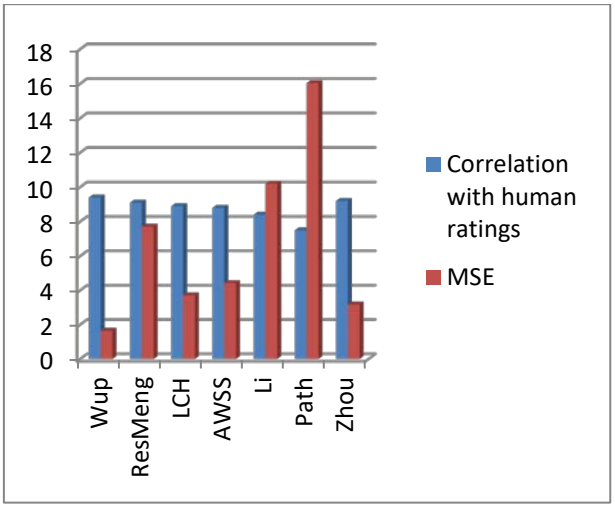


Figure 4: The correlation and MSE values for all measures

6. Conclusion

This research has studied the possibility of applying the traditional semantic similarity measures over AWN. These measures have been applied using Arabic benchmark dataset. The AWN provides information sources which

are: distances, depths and information content of concepts. Therefore, these information sources could be used by different categories of measures such as path-based measures, corpus-dependent information content based measures, and hybrid measures to calculate the similarity score between Arabic word pairs. The AWN has missing information sources such as glosses of concepts. However, some of feature-based measures need these glosses to be applied on AWN. Therefore, Lesk's measure which is well known feature-based measure is not applicable on AWN. Furthermore, the corpus-dependent information content-based measures cannot be applied over AWN due to the ambiguity and sparse data problem. However, to avoid these problems, this research recommends using corpus-independent information content-based measures. The experimental results of applying the traditional semantic similarity measures on AWN found out that Wup measure has the highest correlation value with human ratings. Furthermore, Wup measure has obtained the lowest MSE value against other measures; therefore, this result indicates that the Wup measure has the best performance compared to other measures. Path measure has the worst performance, with lowest correlation with human rating and lowest MSE value.

References

[1] Abderrahim, M. A., Abderrahim, M. E. A., & Chikh, M. A. (2013). Using Arabic wordnet for semantic indexation in information retrieval system. arXiv preprint arXiv:1306.2499.

[2] Abouenour, L., Rosso, P., & Bouzoubaa, K. (2012). IDRAAQ: New Arabic question answering system based on query expansion and passage retrieval.

[3] Al-Khiaty, M. A. R., & Ahmed, M. (2016). UML Class Diagrams: Similarity Aspects and Matching. Lecture Notes on Software Engineering, 4(1).

[4] Slimani, T. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. International Journal of Computer Applications, 80(10), 25-33. http://dx.doi.org/10.5120/13897-1851

[5] Elberrichi, Z., & Abidi, K. (2012). Arabic text categorization: a comparative study of different representation modes. Int. Arab J. Inf. Technol., 9(5), 465-470.

[6] Attia, M. A. (2008). Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation (Doctoral dissertation, University of Manchester).

[7] Fazza, A., James, D., Zuhair, A., & Keeley, A. (2012). Arabic Word Semantic Similarity. Proceedings of World Academy of Science, Engineering and Technology. No. 70. World Academy of Science, Engineering and Technology

[8] Miller, G., & Fellbaum, C. (1998). Wordnet: An electronic lexical database.

[9] Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology, 6(1), 1-12.

[10] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. International journal of lexicography, 3(4), 235-244.

[11] Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006, January). Introducing the Arabic wordnet project. In Proceedings of the third international WordNet conference (pp. 295-300).

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

[13] Cavalli-Sforza, V., Saddiki, H., Bouzoubaa, K., Abouenour, L., Maamouri, M., & Goshey, E. (2013, May). Bootstrapping a WordNet for an Arabic dialect from other WordNets and dictionary resources. In AICCSA (pp. 1-8).

[14] Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In Proceedings of the Annual Meeting of the Association for Computational Linguistics

[15] Michelizzi, J. (2005). Semantic relatedness applied to all words sense disambiguation (Doctoral dissertation, University of Minnesota).

[16] Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

[17] Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. Knowledge and Data Engineering, IEEE Transactions on, 15(4), 871-882.

[18] Almarsoomi, F. A., O'Shea, J. D., Bandar, Z., & Crockett, K. (2013, October). AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 504-509). IEEE.

[19] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 448–453.

[20] Meng, L., Gu, J., & Zhou, Z. (2012). A new model of information content based on concept's topology for measuring semantic similarity in WordNet. International Journal of Grid and Distributed Computing, 5(3), 81-94.

[21] Zhou, Z., Wang, Y., & Gu, J. (2008, November). New model of semantic similarity measuring in wordnet. In Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on (Vol. 1, pp. 256-261). IEEE.

[22] Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. Communications of the ACM, 8(10), 627-633.