World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:6, No:10, 2012

Arabic Word Semantic Similarity

Faaza A, Almarsoomi, James D, O'Shea, Zuhair A, Bandar andKeeley A, Crockett

***Abstract*—**This paper is concerned with the production of an Arabic word semantic similarity benchmark dataset. It is the first of its kind for Arabic which was particularly developed to assess the accuracy of word semantic similarity measurements. Semantic similarity is an essential component to numerous applications in fields such as natural language processing, artificial intelligence, linguistics, and psychology. Most of the reported work has been done for English. To the best of our knowledge, there is no word similarity measure developed specifically for Arabic. In this paper, an Arabic benchmark dataset of 70 word pairs is presented. New methods and best possible available techniques have been used in this study to produce the Arabic dataset. This includes selecting and creating materials, collecting human ratings from a representative sample of participants, and calculating the overall ratings. This dataset will make a substantial contribution to future work in the field of Arabic WSS and hopefully it will be considered as a reference basis from which to evaluate and compare different methodologies in the field.

***Keywords*—**Arabic categories, benchmark dataset, semantic similarity, word pair, stimulus Arabic words

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

I. Introduction

WORD semantic similarity (WSS) has grown to be an important part of natural language processing and information retrieval (IR) for many years. Semantic similarity is an essential component of numerous applications in the fields of artificial intelligence, psychology and computational linguistics, both in the academic community and industry. Examples comprise word sense disambiguation [1], IR [2], semantic search (to find pictures, documents, jobs and videos) [3], [4] and also in the seeking of biological macromolecules such as proteins and DNA [5].

Recently new measures have been proposed to calculate the semantic similarity between two short texts (STSS) of sentence length which rely largely on computing the similarity between words in both sentences [6]. These measures are promising techniques which can play a crucial role in the development of large number of applications. For example, in web page retrieval, STSS measure is used to improve retrieval effectiveness through the calculation of the similarities of page titles [7]. Text mining can also benefit from the use of STSS measure as a criterion to detect unseen knowledge from textual databases [8]. In the conversational agent / dialogue system, the employment of the STSS measure can greatly reduce the scripting process through the use of natural sentences instead of structural patterns of sentences [9].

These applications show that the calculation of semantic similarity between two words is a fundamental task which is frequently represented by similarity between concepts associated with the compared words.

F. Almarsoomi, J.D. O'Shea, Z. Bandar, and K. Crockett are with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M1 5GD, UK. (e-mail: faaza-abdul.j.al-marsoomi@stu.mmu.ac.uk) (e-mail: {j.d.oshea, z.bandar, k.crockett}@mmu.ac.uk).

There are a number of WSS measures [10] in the literature which have been evaluated through the use of the word similarity benchmark dataset before they are integrated into the complete system. Consistency of a WSS measure with human similarity ratings is employed to determine the quality of such measures. This is measured as the product-moment correlation coefficient computed between the set of human similarity ratings and those from the word similarity measure using a benchmark dataset [11].

To date, most of the reported word similarity measures are for English. However, there is no work done specifically for the Arabic language. Consequently, there is no Arabic word semantic similarity dataset. In order to improve the accuracy of a large number of Arabic applications [12], [13], it is important first to create an Arabic word semantic similarity dataset using the best possible available methods which will make a substantial contribution to future work in the field of Arabic WSS.

The focus of this paper is the production of the first word similarity benchmark dataset for Modern Standard Arabic (MSA) which is the formal language of the Arab world. Arabic is a Semitic language which is spoken by over 330 million people [14]. The Arabic alphabet uses 25 consonants and 3 long vowels which are written from right to left. These letters take different shapes 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 [15]. Also Arabic words exhibit a complex internal structure, where words often incorporate affixes that mark grammatical inflections and clitics to signify different parts of speech [15].

In this paper, the first Arabic word similarity dataset is created which consists of 70 Arabic word pairs with human ratings. The methodology comprises of four fundamental steps which includes materials be gathered (word pairs), human ratings collected, overall ratings computed and the dataset validated. This methodology is described and illustrated in this paper.

The remaining sections of this paper are organized as follows: section 2 reviews the prior work on word semantic similarity measures and datasets. Section 3 describes the procedure of the production of the Arabic dataset which includes constructing the set of Arabic word pairs experiment and collecting human ratings experiment. Section 4 discusses the experimental results and compares the Arabic dataset with related work.

II. Prior work

A number of algorithms have been developed for measuring WSS; most of these measures are for the English language. The following sections provide a brief review of existing WSS measurements and the datasets used for comparing and evaluating them.

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

A. Word Semantic Similarity Measure

Existing WSS measures can be generally categorized into three groups based on the information source they exploit: Dictionary / Ontology based methods [16], [17] typically use the semantic information derived from knowledge bases to compute the WSS. Corpus-based methods [18] principally use the frequency of a word's occurrence to calculate WSS using statistical information derived from the large corpora. Hybrid methods [10], [19] calculate the WSS by combining multiple information sources. A detailed review of WSS measures can be obtained in [20], [21].

B. Word Similarity Benchmark Dataset

WSS measures have been evaluated using the word similarity benchmark dataset before they are integrated into the complete system. Two word benchmark datasets are commonly used for evaluating and comparing new developments, both of them for English language.

Rubenstein & Goodenough R&G [22] created the most influential word benchmark dataset for English. The procedure of the production of this dataset comprised of two steps. The first step involved generating 65 word pairs ranging from maximum to minimum similarity of meaning. A list of 48 English nouns represented in two columns (A and B) was employed to produce the 65 word pairs by selecting one word from column A and one from column B. The second step involved collecting the human similarity ratings of the 65 word pairs. 51 undergraduate participants were asked to assess the similarity between the word pairs based on how similar they were in meaning. The words pairs were ranked using a rating scale which ran from 0 (minimum similarity) to 4 (maximum similarity). However R&G dataset was published without justification for the specific choices of 48 nouns and the method of the combination of word pairs.

Miller & Charles (M&C) [23] replicated the R&G experiment and considered only 30 word pairs from the 65 word pairs of the R&G dataset to avoid an inherent bias towards low similarity. 38 undergraduate students (all Native English speakers) were asked to rank the 30 word pairs using a rating scale from 0 to 4. This experiment was performed 25 years after the R&G experiment, however the correlation between human ratings in the two datasets obtained a high value of 0.97. The M&C experiment was replicated by Resink [11] in 1995. The subset of 30 word pairs was ranked by the sample of 10 computer science graduate students and post-docs. This experiment obtained a high value correlation of 0.96 with M&C dataset. The results of these experiments show that the R&G dataset has indicated stability over the years. This stability illustrates that the use of human ratings could be a reliable reference for the purpose of comparison with computational methods.

The R&G dataset is still valuable 45 years after it was produced [21]. Therefore the R&G methodology is used as a general framework to produce the first word benchmark dataset for Arabic.

III. Production of the Arabic Word Similarity Benchmark Dataset

The methodology of the production of the Arabic dataset involved conducting two experiments. The aim of experiment 1 was to construct the set of Arabic word pairs, whilst the aim of experiment 2 was to collect the human similarity ratings.

Furthermore, five fundamental hurdles were taken into consideration as a part of the Arabic word dataset design process:

1) Selecting a sample of participants representing the general human population. Because the dataset was created for Arabic, it was decided to use a representative sample of participants from different Arabic countries which signify the general population taking into account the subject knowledge, gender, and age.
2) Representation of the Arabic language with a delimited number of word pairs. A new method (described in section III.*A*) was used to select the stimulus Arabic words. These words were selected and presented in a way that contributes to the control of the range of semantic similarity (maximum to minimum) covered by the set of produced word pairs.
3) Selecting a representative sample of Arabic word pairs. This was achieved by conducting an experiment to generate the set of Arabic word pairs using human judgments.
4) Selecting the measurement scale. The type of statistical methods that can be applied to the similarity measures is defined based on the measurement scale used when they created. A ratio scale was used as a measurement scale in the prior work for both WSS measures and word similarity dataset [11], [22], and [23]. This dataset is intended to assess the accuracy of the algorithms (WSS) running on the scale from 0 (minimum similarity) to maximum which is a kind of ratio scale.
5) Collection of the ratings that precisely signify human conception of similarity. A combination of card sorting and semantic anchors (described in section III.*C*) was used as the most suitable procedure to collect human similarity ratings. This combination was selected based upon four experiments [24] which examined the impact of varying two factors, Order (randomize the order of the word pairs) and Anchors, on human ratings. The experimental results showed that one of the combinations, known as Card Sorting with Semantic Anchors was superior as it obtained significantly lower noise and a higher correlation coefficient.

A. Selecting the Set of Stimulus Arabic Words

The first step of the production of the Arabic dataset was to create a list of Arabic words which was presented later to produce the set of Arabic word pairs using human judgments. The decision was made to use categories known as category norms to select stimulus words for producing a list of Arabic words.

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

A category norm is defined as a set of words within the same theme, listed by frequency, which is created as responses by human participants to a specific category [25]. These categories consist of a large number of different themes used in many studies. For example, English category norms consist of 56 to 70 different themes used in 1600 projects after they were produced [26]. It was decided to employ category norms for selecting the set of stimulus words based on the two important features of these categories (a large number of different themes and a list of words within the same theme).

Due to the lack of category norms for the Arabic language, 27 Arabic categories were created and employed to select the stimulus Arabic words. As in category norms, the Arabic categories have different themes and consist of ordinary Arabic words. The words in each category are more similar to each other than to the words of other categories. The following steps illustrate the production of Arabic categories:

Step1. 22 categories were created to have the same themes as R&G to take advantage of four decades of experience with this dataset. The list of English words in the R&G experiment contains 48 nouns (24 pairs) for 22 different themes. This list was employed to create the 22 Arabic categories consisting of 22 different themes as follows:

1) For each English pair, the two nouns were translated into Arabic using the first meaning from an established English–Arabic dictionary [27]. To ensure translation accuracy, the translated nouns were checked by a professional translator and a lecturer fluent in both languages.
2) Based on the definition of two selected nouns [28], the Arabic category was given a specific name and a set of Arabic nouns (described in one word) within the same category theme were added for the production of the entire category.

For example, the English nouns (Gem and Jewel) were selected (same theme) and both were translated into (جوهرة) in Arabic. The Arabic category was created and called the Gemstones category (احجار كريمة) based on the definitions of jewel (*a precious stone used to decorate valuable things that you wear, such as rings or necklaces*) and gem (*a jewel or stone that is used in jewelry*). A set of Arabic words within the same category theme (Diamond /ماس, Pearl /لؤلؤ, Crystal /بلور, ...) were added to produce an entire category.

Some English nouns were omitted and not added to Arabic categories due to translation problems. First, some English nouns translated into the same Arabic word such as (*Gem and Jewel*) both translated as جوهرة in Arabic. Also some English nouns were translated into two Arabic words such as the English noun *Madhouse* in Arabic translates as مستشفى المجانين. Consequently, all translated nouns (described in two words or having the same translated word) were omitted and not added to the Arabic categories. Table I illustrates the English nouns and the reasons of omission.

As a result, 22 Arabic categories were produced from 48 translated nouns as shown in Table II.

Step2. 5 new categories were created to expand the 22 categories' themes and incorporate particular Arabic themes as shown in Table II.

For example, the Arabic categories created in the first step have the type of male life stages category, to expand this theme and include male and female, the type of female life stages category was created. Religious events and type of lifestyle categories were produced to incorporate particular Arabic themes.

Using the Arabic categories created in step 1 and 2, the first two nouns were selected from each category to generate the set of 56 stimulus Arabic words which consisted of 27 different themes as shown in Table III.

TABLE I
ENGLISH NOUNS WITH THE REASONS OF OMISSION

	English Nouns	Arabic Nouns	The reason of omitting
1	Madhouse	مستشفى المجانين	Described in two words
2	Asylum	مستشفى المختلين	Described in two words
3	Gem / Jewel	جوهـره	Same translated word
4	Sage / Oracle	حكيـم	Same translated word
5	Slave / Serf	عبـد	Same translated word
6	Tool / Implement	اداة	Same translated word
7	Hill / Mound	تـل	Same translated word
8	Car / Automobile	سيـارة	Same translated word
9	Cock / Rooster	ديـك	Same translated word
10	Graveyard/ Cemetery	مقبـرة	Same translated word

TABLE II
THE LIST OF ARABIC CATEGORIES

	Categories Names	اسماء الفئات العربية
1	Medical Places	مواقع طبية
2	Handwritten text	نص مكتوب يدويا
3	Type of male's life stages	مراحل حياة الذكر
4	Member of the clergy	رجل دين
5	Transportation vehicles	مركبات نقل
6	Coastal area	منطقة ساحلية
7	Bird	طير
8	Type of furnishings	نوع من المفروشات
9	Source of a human body energy	مصدر طاقة جسم الانسان
10	Appliance for cooking	جهاز طهي
11	Gemstones	أحجار كريمة
12	Drinking utensil	ادوات او آنية للشرب
13	Geographic	جغرافية الارض
14	Parts of day	اجزاء اليوم
15	Type of equipment	نوع من معدات/ تجهيزات
16	Type of departure	نوع من رحيل/ مغادرة
17	Somebody practices witchcraft	شخص يمارس السحر
18	Wise person	شخص حكيم
19	Facial expressions	تعابير الوجهة
20	Material for tying things	مادة لربط الاشياء
21	Person in slavery	شخص في العبودية
22	Burial place	اماكن لدفن الاموات
23	Religious events	احداث دينية
24	Type of lifestyle	نوع من نمط / اسلوب الحياة
25	Type of female life stages	مراحل حياة الانثى
26	Vacation activities	انشطة العطلات
27	Family members	أعضاء العائلة

B. Experiment 1: Construction of the Set of Arabic Word Pairs

1. Participants

A sample of 22 Arabic native speakers was chosen to perform the task of generating the set of Arabic word pairs. The participants were from different Arabic countries which include: Iraq, Saudi Arabia, Jordan, Libya, and Palestine. The sample consisted of 10 academics (University lecturers) and 12 non-academics.

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

They were 13 Science/Engineering vs. 9 Art/Humanities backgrounds. The average age was 34 years and the standard deviation (SD) was 6.3 with 13 female and 9 male.

2. Materials

A list of Arabic nouns was created through the use of the set of stimulus Arabic words (selected in section III.*A*). This was done by representing the set of 56 stimulus words in two columns (A and B) with each column containing 28 different Arabic words.

As shown in Table III the list of Arabic nouns consists of 28 pairs of nouns and the nouns of each pair within the same theme such as *Hospital* and *Infirmary* (one noun (*Hospital*) in column A and one (*Infirmary*) in column B).

The order of Arabic nouns in column B was randomized to minimize ordering effects. This list was presented to 22 Arabic participants to generate the set of Arabic word pairs ranging from high to low similarity of meaning.

Two recording sheets were used by 22 Arabic participants containing instructions (described in section *B.3*) to create two lists of Arabic word pairs which included: a High Similarity of Meaning list (HSM) containing 28 word pairs between strongly related and identical in meaning.

A Medium Similarity of Meaning list (MSM) containing 32 word pairs between vaguely similar and very much alike in meaning while a low similarity of meaning list was selected randomly.

Because the list of Arabic nouns has 28 noun pairs (each pair has the same theme), the participants were requested to write 28 high similarity word pairs. Unlike the high and low similarity word pairs, it is relatively difficult for humans to write medium similarity word pairs. So, to increase the opportunity of obtaining medium similarity word pairs, the participants were asked to write 32 word pairs for (MSM) list.

3. Procedure

The list of Arabic nouns was employed to produce the set of Arabic word pairs by selecting one word from column A and one from column B based on the amount of similarity of meaning.

The participants were instructed to perform the following task.

1) Using the list of Arabic nouns, write a list of 28 Arabic word pairs that have HSM.
2) The Arabic word pairs always contain one word from column A and one from column B.
3) The HSM list contains word pairs between strongly related and identical in meaning.
4) Please write 28 word pairs because all uncompleted questionnaires must be ignored.

Following the same procedure, the participants were requested to write a list of 32 Arabic word pairs for MSM.

Some notes were included in the instruction sheet which stated: "You can select any word from column A more than once with different words from column B to create new word pairs"; and also "Please do not write the same word pair more than once in the same sheet or between different sheets".

TABLE III
THE LIST OF ARABIC NOUNS

	Column A			Column B	
1	Hospital	مستشفى	1	Bus	باص
2	Signature	توقيع	2	Pigeon	حمامة
3	Boy	صبي	3	Grave	قبر
4	Master	سيد	4	Woodland	أحراش
5	Coach	حافلة	5	Vegetable	خضار
6	Coast	ساحل	6	Mountain	جبل
7	Hen	دجاجة	7	Means (noun)	وسيلة
8	Cushion	مسند	8	Diamond	الماس
9	Food	طعام	9	Travel (noun)	سفر
10	Stove	موقد	10	Lad	فتى
11	Gem	جوهرة	11	Infirmary	مشفى
12	Glass	كأس	12	Magician	مشعوذ
13	Forest	غابة	13	Midday	ظهيرة
14	Hill	تل	14	Sheikh	شيخ
15	Noon	ظهر	15	Pillow	مخدة
16	Tool	اداة	16	Thinker	مفكر
17	Journey	رحلة	17	Odalisque	جارية
18	Wizard	ساحر	18	Shore	شاطئ
19	Sage	حكيم	19	Endorsement	تصديق
20	Smile	ابتسامة	20	Laugh	ضحك
21	Cord	حبل	21	Oven	فرن
22	Slave	عبد	22	String	خيط
23	Sepulcher	ضريح	23	Tumbler	قدح
24	Feast	عيد	24	Young woman	شابة
25	Countryside	ريف	25	Walk (noun)	مشي
26	Run (noun)	جري	26	Sister	أخت
27	Brother	أخ	27	Fasting	صيام
28	Girl	فتاة	28	village	قرية

4. Experimental Results

A set of 70 Arabic word pairs were selected using the two lists of word pairs (HSM and MSM lists) generated through experiment 1 plus the list of low similarity word pairs which were selected randomly. Table IV illustrates the final set of Arabic word pairs, where the first and last columns represent the set of Arabic word pairs in English and Arabic. The second column contains the number of participants who chose the word pair.

1) The first 24 word pairs in table IV represent the high similarity word pairs which were selected using HSM list. Those word pairs were chosen by all the 22 participants.
2) The word pairs from 25 to 47 (23 pairs) represent the medium similarity word pairs which were chosen by more than half the participants.
3) The last 23 word pairs were selected to represent the low similarity word pairs. A combination of medium similarity candidate word pairs rated low by participants plus randomly selected low similarity word pairs (using the list of Arabic nouns) to allow for word pairs that were not chosen by the participants.

For each noun in the list of Arabic nouns, the frequency of appearance of this noun in the final set of Arabic word pairs was calculated. The nouns which have an occurrence of more than two times were removed from the list of Arabic nouns to avoid a biased set of nouns from being used. The remaining Arabic nouns were used to generate a list of Arabic word pairs randomly. High and medium similarity word pairs already found by participants were removed. The remaining pairs were selected at random as they were good candidates for low similarity.

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

C. Experiment 2: Collection the Human Similarity Ratings

1. Participants

60 participants from different Arabic countries were asked to rank the set of 70 Arabic word pairs collected in Experiment 1. All were Arabic native speakers who had not taken part in Experiment 1 and they were from 7 Arabic countries which included: Iraq, Saudi Arabia, Egypt, Jordan, Kuwait, Libya, and Palestine. The participants were equally balanced between students and non-students which they were 39 Science/Engineering vs. 21 Art/Humanities backgrounds. The average age was 29 years and the standard deviation (SD) was 7.2 with an equal balance of male and female.

2. Materials

The set of 70 Arabic word pairs collected in experiment 1 were presented to Arabic participants to collect judgments on how similar they are in meaning. Each of 70 word pairs was printed on a separate card. Each participant was given an envelope containing 70 cards (the order of 70 cards was initially randomized to minimize the ordering effects) and 3 sheets which included: instructions for collecting the human rating, a similarity rating recording sheet and a personal information sheet.

TABLE IV
THE FINAL SET OF ARABIC WORD PAIRS

	Word Pairs		Participants	أزواج الكلمـات	
1	Boy	Lad	22	صبي	فتى
2	Coast	Shore	22	ساحل	شاطئ
3	Cushion	Pillow	22	مسند	مخدة
4	Gem	Diamond	22	جوهرة	الماس
5	Glass	Tumbler	22	كأس	قدح
6	Forest	Woodland	22	غابة	أحراش
7	Noon	Midday	22	ظهر	ظهيرة
8	Tool	Means	22	اداة	وسيلة
9	Journey	Travel	22	رحلة	سفر
10	Smile	Laugh	22	إبتسامة	ضحك
11	Countryside	Village	22	ريف	قرية
12	Girl	Young woman	22	فتاة	شابة
13	Signature	Endorsement	22	توقيع	تصديق
14	Coach	Bus	22	حافلة	باص
15	Hen	Pigeon	22	دجاجة	حمامة
16	Sepulcher	Grave	22	ضريح	قبر
17	Run	Walk	22	جري	مشي
18	Hospital	Infirmary	22	مستشفى	مشفى
19	Master	Sheikh	22	سيد	شيخ
20	Wizard	Magician	22	ساحر	مشعوذ
21	Feast	Fasting	22	عيد	صيام
22	Food	Vegetable	22	طعام	خضار
23	Stove	Oven	22	موقد	فرن
24	Hill	Mountain	22	تل	جبل
25	Sage	Thinker	21	حكيم	مفكر
26	Cord	String	21	حبل	خيط
27	Slave	Odalisque	21	عبد	جارية
28	Brother	Sister	21	أخ	أخت
29	Hen	Oven	20	دجاجة	فرن
30	Coach	Means	19	حافلة	وسيلة
31	Sage	Sheikh	18	حكيم	شيخ
32	Girl	Sister	16	فتاة	أخت
33	Journey	Shore	15	رحلة	شاطئ
34	Coast	Mountain	14	ساحل	جبل
35	Master	Thinker	14	سيد	مفكر
36	Coach	Travel	14	حافلة	سفر
37	Food	Oven	14	طعام	فرن
38	Brother	Lad	13	أخ	فتى
39	Girl	Odalisque	13	فتاة	جارية
40	Slave	Lad	13	عبد	فتى
41	Feast	Laugh	13	عيد	ضحك
42	Hospital	Grave	12	مستشفى	قبر
43	Hill	Woodland	12	تل	أحراش
44	Journey	Bus	12	رحلة	باص
45	Tool	Tumbler	12	اداة	قدح
46	Run	Shore	11	جري	شاطئ
47	Tool	Pillow	11	اداة	مخدة
48	Sepulcher	Sheikh	10	ضريح	شيخ
49	Cord	Mountain	9	حبل	جبل
50	Gem	Young woman	8	جوهرة	شابة
51	Countryside	Vegetable	7	ريف	خضار
52	Glass	Fasting	6	كأس	صيام
53	Forest	Shore	5	غابة	شاطئ
54	Noon	Fasting	4	ظهر	صيام
55	Glass	Diamond	3	كأس	الماس
56	Signature	String	2	توقيع	خيط
57	Boy	Midday	1	صبي	ظهيرة
58	Wizard	Infirmary	0	ساحر	مشفى
59	Cushion	Diamond	0	مسند	الماس
60	Noon	String	0	ظهر	خيط
61	Boy	Endorsement	0	صبي	تصديق
62	Gem	Pillow	0	جوهرة	مخدة
63	Cord	Midday	0	حبل	ظهيرة
64	Countryside	Laugh	0	ريف	ضحك
65	Hill	Pigeon	0	تل	حمامة
66	Slave	Vegetable	0	عبد	خضار
67	Smile	Village	0	إبتسامة	قرية
68	Stove	Walk	0	موقد	مشي
69	Coast	Endorsement	0	ساحل	تصديق
70	Smile	Pigeon	0	إبتسامة	حمامة

3. Procedure

A combination of card sorting (sorting the cards based on the amount of similarity of meaning) and semantic anchors were used in this experiment to collect human judgments. A semantic anchor permits the participants to map a scale descriptor to each of the major scale points [24]. 5 semantic anchors for the 5 point rating scale listed in Table V were used in this experiment.

The participants were requested to rate each word pair based on how similar they were in meaning after sorting the cards. Also they ranked each word pair using the 5 points rating scales which ran from 0.0 (unrelated in meaning) to 4.0 (identical in meaning).

The participants were asked to perform the following task:

1) Please sort the 70 cards into four groups according to the similarity of meaning. The HSM group contains word

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

pairs between strongly related and identical in meaning, the two MSM groups contain word pairs vaguely similar or very much alike in meaning and low similarity contains word pairs unrelated in meaning.

2) The number of cards in each group is based on your judgment of each card.
3) Please check the cards in each group carefully; you may change a word pair from group to another at this stage.
4) Please rate each word pair according to the similarity of meaning using the rating scale points.

Furthermore, some notes were included in the instruction sheet which stated: "Please do not write values greater than 4.0 or less than 0.0. Also, you may rate more than one pair with the same value." And: "You can use the first decimal place to assign an accurate degree of similarity (for instance, if you think the similarity of word pair is between 2 and 3 you can assign a value such as 2.5)".

TABLE V
SEMANTIC ANCHORS

Rating scale	Semantic Anchors	
0	The word pairs are unrelated in meaning	زوج الكلمات لا يوجد ارتباط بينها في المعنى
1	The word pairs are vaguely similar in meaning.	زوج الكلمات بينها تشابه ضمني في المعنى
2	The word pairs are very much alike in meaning.	زوج الكلمات التي بينها تشابه واضح (اكثر من ضمني)
3	The word pairs are strongly related in meaning	زوج الكلمات التي بينها علاقة قوية في المعنى
4	The word pairs are identical in meaning	زوج الكلمات المترادفة او المتطابقة في المعنى

4. Experimental Results

Table VI contains the result of experiment 2 which represents the set of Arabic word pairs with a human similarity rating. The first and last pairs of columns represent the set of Arabic word pairs in English and Arabic. The third column contains the average of similarity rating collected from 60 Arabic native speakers.

Fig. 1 shows the correlation coefficients of 60 participants, where the consistency of similarity rating for each participant with the rest of group was determined using the Pearson product moment correlation coefficient. This was calculated by the leave-one-out resampling technique [11] for the ratings of each participant with all of the rest of the group.

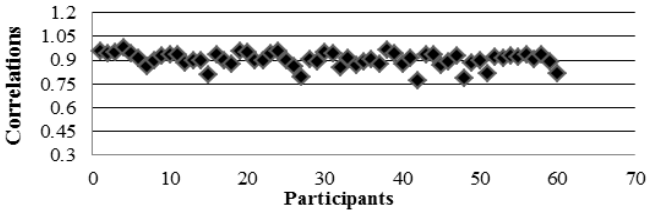

Fig. 1 Correlation coefficients of 60 participants

IV. DISCUSSION

A. The Benchmark Dataset

The human similarity ratings collected in experiment 2 are calculated as the mean of the judgments provided by the 60 participants for each of the Arabic word pairs as shown in Table VI.

The correlation coefficient is considered as a suitable measure for consistency [24]. The consistency between the set of human ratings and those obtained from the WSS algorithms is determined using the Pearson product-moment correlation coefficient which is considered suitable for measures created on a ratio scale [24].

The average of the correlations of all participants on the Arabic dataset was calculated; this can be used to assess the performance of a computational (WSS) attempt to carry out the same task. Any WSS measure which equals or exceeds the average of the correlations of all participants is considered to be performing well. As shown in Table VII, the average of the correlations of all participants for the Arabic dataset is 0.902. The worst performing participant of 0.767 is considered as the lower bound for the expected performance whereas any machine measure coming close to the best performing participant at 0.974 would be considered as performing very well.

TABLE VII
PEARSON CORRELATION COEFFICIENT WITH MEAN HUMAN JUDGMENTS

	Correlation r
Average of the correlation of all participants	0.902
Best participant	0.974
Worst participant	0.767

Both high similarity and low similarity word pairs are subject to very consistent human judgments, as shown in Fig. 2 and Fig. 3. Unlike the low and high similarity word pairs, the human ratings of the medium similarity word pairs spread more evenly across the similarity range (0 to 4). Consequently, the medium similarity word pairs have higher values of SD than the other word pairs.

TABLE VI
THE SET OF ARABIC WORD PAIRS WITH HUMAN RATINGS

	Word Pairs		Human Ratings	أزواج الكلمـات			Word Pairs		Human Ratings	أزواج الكلمـات	
1	Coast	Endorsement	0.03	تصديق	ساحـل	36	Slave	Lad	1.77	فتى	عبد
2	Noon	String	0.03	خيط	ظهر	37	Journey	Bus	1.83	باص	رحلة

3	Cushion	Diamond	0.06	الماس	مسند
4	Gem	Pillow	0.07	مخدة	جوهرة
5	Stove	Walk	0.07	مشي	موقـد
6	Cord	Midday	0.08	ظهيرة	حبـل
7	Signature	String	0.08	خيط	توقيـع
8	Boy	Endorsement	0.12	تصديـق	صبـي
9	Boy	Midday	0.16	ظهيرة	صبـي
10	Slave	Vegetable	0.16	خضار	عبـد
11	Smile	Village	0.18	قريـة	إبتسامة
12	Smile	Pigeon	0.20	حمامة	إبتسامة
13	Wizard	Infirmary	0.22	مشفى	ساحر
14	Noon	Fasting	0.29	صيـام	ظهـر
15	Hill	Pigeon	0.33	حمامة	تـل
16	Countryside	Laugh	0.34	ضحك	ريف
17	Glass	Diamond	0.36	الماس	كـأس
18	Glass	Fasting	0.38	صيـام	كـأس
19	Cord	Mountain	0.54	جبـل	حبـل
20	Hospital	Grave	0.83	قبر	مستشفى
21	Forest	Shore	0.86	شاطئ	غابـة
22	Gem	Young woman	0.87	شابة	جوهرة
23	Sepulcher	Sheikh	0.89	شيخ	ضريـح
24	Tool	Pillow	0.99	مخدة	اداة
25	Coast	Mountain	1.06	جبل	ساحل
26	Run	Shore	1.13	شاطئ	جري
27	Hill	Woodland	1.19	أحراش	تـل
28	Countryside	Vegetable	1.24	خضار	ريـف
29	Tool	Tumbler	1.32	قدح	اداة
30	Master	Thinker	1.36	مفكر	سيـد
31	Feast	Laugh	1.36	ضحك	عيـد
32	Hen	Oven	1.44	فرن	دجاجـة
33	Journey	Shore	1.47	شاطئ	رحلـة
34	Coach	Travel	1.60	سفر	حافلـة
35	Food	Oven	1.76	فرن	طعام
38	Girl	Odalisque	1.96	جاريـة	فتـاة
39	Feast	Fasting	1.96	صيـام	عيـد
40	Coach	Means	2.07	وسيلـة	حافلـة
41	Brother	Lad	2.15	فتـى	أخ
42	Sage	Sheikh	2.26	شيـخ	حكيم
43	Girl	Sister	2.38	أخت	فتـاة
44	Hill	Mountain	2.60	جبل	تـل
45	Hen	Pigeon	2.61	حمامة	دجاجة
46	Master	Sheikh	2.66	شيـخ	سيـد
47	Food	Vegetable	2.78	خضار	طعـام
48	Slave	Odalisque	2.84	جاريـة	عبـد
49	Run	Walk	3.01	مشي	جـري
50	Brother	Sister	3.08	أخت	أخ
51	Cord	String	3.09	خيـط	حبـل
52	Forest	Woodland	3.14	أحراش	غابـة
53	Sage	Thinker	3.30	مفكر	حكيم
54	Gem	Diamond	3.38	الماس	جوهرة
55	Cushion	Pillow	3.38	مخدة	مسند
56	Journey	Travel	3.39	سفر	رحلـة
57	Countryside	Village	3.41	قريـة	ريـف
58	Smile	Laugh	3.48	ضحـك	إبتسامة
59	Stove	Oven	3.55	فـرن	موقـد
60	Coast	Shore	3.56	شاطئ	ساحـل
61	Signature	Endorsement	3.58	تصديـق	توقيـع
62	Tool	Means	3.68	وسيلـة	اداة
63	Noon	Midday	3.70	ظهيـرة	ظهـر
64	Boy	Lad	3.71	فتـى	صبـي
65	Girl	Young woman	3.74	شابـة	فتـاة
66	Sepulcher	Grave	3.75	قبـر	ضريـح
67	Wizard	Magician	3.76	مشعوذ	ساحـر
68	Coach	Bus	3.80	بـاص	حافلـة
69	Glass	Tumbler	3.82	قدح	كـأس
70	Hospital	Infirmary	3.91	مشفى	مستشفى

For example, the word pair 46 (سيد شيخ) has SD 1.07 and the mean of human ratings 2.66. The distribution of the human ratings for this word pair should be grouped around a peak 2.66. In fact the module class is 3 and the distribution is relatively flat as shown in Fig. 4.

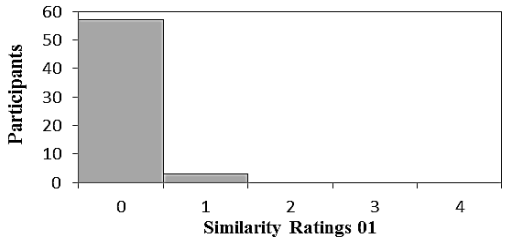

Fig. 2 Histogram of similarity ratings for word pair 01, SD=0.14

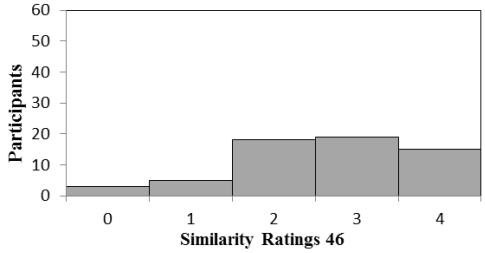

Fig. 3 Histogram of similarity ratings for word pair 70, SD=0.28

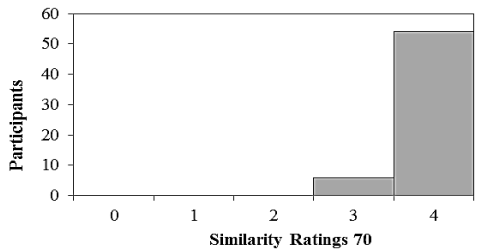

Fig. 4 Histogram of similarity ratings for word pair 46, SD=1.07

B. A Comparison with the R&G Dataset

The most influential word dataset for English to R&G was used as a general framework for the production of the Arabic word dataset. In this section, a comparison is conducted between the two datasets to illustrate the differences between them.

1. Method of Selection of Materials

48 nouns (22 themes) to the R&G dataset were employed to make up the set of 65 word pairs in a variety of combinations which covered a range of semantic similarity values from high to low.

However, the R&G dataset was published without justification for the specific choices of 48 nouns and the method of the combination of word pairs. The R&G dataset is skewed towards low similarity word pairs [23].

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864

For this study 56 stimulus Arabic words (27 themes) were carefully selected through the use of 27 Arabic categories to generate the set of 70 Arabic word pairs. Semantic similarity judgments are an issue of human perception. Experiment 1 was used to create 70 word pairs spanning the similarity range based on human judgments to counter the bias towards low similarity in the R&G dataset.

2. Sampling the Population of Participants

The sample of participants used in the R&G experiment to collect human ratings was two groups of college undergraduates for a total of 51 participants. No information was provided on the composition of age or gender for each group and whether the sample of participants used in this experiment contained only native English speakers.

The sample of human population used in the Arabic dataset experiments is more representative than the R&G experiment. The value of a sample of participants selected to carry out a specific experiment could be reduced as a representative sample if there is a high homogeneity of participants and they are distant from the general population [24].

Consequently, the sample of Arabic participants was selected as a general population (students and non-students) from different Arabic countries taking account of the gender, age, and academic background factors. The sample was selected to balance gender (males and females), student and non-student, academic background (science/engineering vs. arts/humanities) and age to avoid a bias towards any element of these factors.

3. The Procedure of Collection Human Ratings

A card sorting technique was used for collecting human ratings in the R&G experiment. The 65 word pairs were presented to collect the human judgments. Each word pair was printed on a separate slip and the order of 65 slips was randomized before presentation. The participants were asked to sort the slips into order of similarity of meaning and each word pair was rated by assigning a value from 4.0- 0.0: the greater the similarity of meaning the higher the number.

A combination of card sorting with semantic anchors was used to collect human ratings in the Arabic dataset experiment, which is considered as the best currently known experimental practice.

Each word pair in the dataset was printed on a separate card and the order of 70 cards was randomized before presentation. The participants were asked to sort the cards into four groups based on the similarity of meaning. The word pairs in each group were rated using a point rating scale (the points described by the semantic anchors) which ran from 0 (low similarity) to 4 (high similarity).

V. Conclusion

This paper has described the production of the first Arabic benchmark dataset for WSS algorithms. Though it is not possible to cover the language comprehensively in this dataset (70 word pairs), a new method was used to select the 56 stimulus Arabic words through the creation of 27 Arabic categories with 27 different themes to promote the best possible semantic representation.

Unlike the prior work [22], participants were chosen to produce 70 word pairs which covered a range of word semantic similarity values from high (e.g. مستشفى - مشفى) to low (e.g. ساحل - تصديق). Human ratings were collected using the best currently known experimental practice and the statistical methods applied to calculate the overall ratings and defined the lower and upper bound for performance were the mean of human judgments and the Pearson Product-Moment correlation coefficient respectively. The sample of participants used in the Arabic dataset experiments were selected to get a balance and representation of the human population well beyond that of prior work. Furthermore, the procedure used for production of this dataset can be used by other Arabic researchers to extend the Arabic WSS benchmark dataset. Unfortunately, there are no WSS measures for Arabic, however the developments in English clearly point out the need for them. Also Arabic researchers are introducing the components required in terms of ontologies and corpora to produce such measures. Therefore, we present this dataset for future development and hopefully this will motivate Arabic researchers to start experimenting with Arabic word semantic similarity dataset. We are currently developing an Arabic word semantic similarity measure for calculating the similarity between concepts associated with the compared words in the Arabic lexical database known as Arabic wordnet [29]. The accuracy of this measure will be assessed using the Arabic word dataset developed in this paper.

References

[1] S. Ravi, and M. Rada, "Unsupervised graph-based word sense disambiguation using measures of word semantic similarity," *In Proceedings of ICSC*, 2007.

[2] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. E. Milios, "Information retrieval by semantic similarity," *International Journal on Semantic Web and Information Systems*, vol. 2, no. 3, pp. 55-73, 2006.

[3] J. Davies, U. Krohn, and R. Weeks, "QuizRDF: Search technology for the semantic web," WWW2002 *workshop on RDF and Semantic Web Applications, 11th International WWW Conference WWW2002*, Hawaii, USA, 2002.

[4] Y. Aytar, M. Shah, and L. Jiebo, "Utilizing semantic word similarity measures for video retrieval," *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR08)*, pp. 1–8, Jun. 2008.

[5] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between Gene ontology terms," *Data & Knowledge Engineering*, vol. 61, no. 1, pp. 137-152, 2007.

[6] H. Chukfong, A. Masrah, M. Azmi, A. Rabiah and C. Shyamala, "Word sense disambiguation based sentence similarity," *Coling 2010: Poster Volume*, pp. 418–426, Beijing, Aug. 2010.

[7] E.K. Park, D.Y. Ra, and M.G. Jang, "Techniques for improving web retrieval effectiveness," *Information Processing and Management*, vol. 41, no. 5, pp. 1207–1223, 2005.

[8] J. Atkinson-Abutridy, C. Mellish, and S. Aitken, "Combining information extraction with genetic algorithms for text mining," *IEEE Intelligent Systems*, vol. 19, no. 3, 2004.

[9] K. O'Shea, Z. Bandar, and K. Crockett, "A Conversational agent framework using semantic analysis," *International Journal of Intelligent Computing Research (IJICR)*, vol. 1, no. 1, Mar. 2010.

[10] V. S. Zuber, and B. Faltings, "OSS: A semantic similarity function based on hierarchical ontologies," *In Proceedings of IJCAI*, pp. 551-556, 2007.

[11] P. Resnik, "Information content to evaluate semantic similarity in a taxonomy," *In Proceedings of IJCAI*, pp. 448-453, 1995.

[12] M. Diab, M. Alkhalifa, S. ElKateb, C. Fellbaum, A. Mansouri, and M. Palmer, "Semeval-2007 task 18: Arabic semantic labelling," *In*

Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 2007.

[13] M. Hijjawi, *ArabChat : an Arabic Conversational Agent*. PhD. Thesis, Department of Computing and Mathematics, Faculty of Science and Engineering, Manchester Metropolitan University, UK, 2011.

[14] A. Farghaly, K. Shaalan, "Arabic natural language processing: challenges and solutions," *ACMTransactions on Asian Language Information Processing*, vol. 8, no. 4, Article 14, 2009.

[15] N. Y. Habash, *Introduction to Arabic Natural Language Processing*. Graeme Hirst 2010. Morgan &Claypool, 2010, PP 11-12 & 39-41.

[16] M. Jarmasz, and S. Szpakowicz, "Roget's Thesaurus and semantic similarity," *In proceedings of the international conference on Recent Advances in Natural Language processing*, Borovetz, Bulgaria, pp. 212-219, 2003.

[17] R. Rada, H. Mili, M. Bicknell, and E. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 19, pp 17-30, 1989.

[18] D. Lin, "An Information-theoretic definition of similarity," *In Proceedings of Conference on Machine Learning*, pp. 296-304, 1998.

[19] Y. Li, Z. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871-882, 2003.

[20] T. Pedersen, V. S. Pakhomov, S. Patwardhan, and C.G. Chute, "Measures of semantic similarity and relatedness in the Biomedical Domain," *Journal of Biomedical Informatics*, vol. 40, PP. 288-299, 2007.

[21] G. Pirro, "Semantic similarity metric combining features and intrinsic information content," *Data & Knowledge Engineering*, vol. 68. pp. 1289-1308, 2009.

[22] H. Rubenstein, and J. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, Vol. 8, pp.627–633, 1965.

[23] G.A. Miller, and W.G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, pp.1–28, 1991.

[24] J.D. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "Benchmarking Short Text Semantic Similarity," *Int. J. Intelligent Information and Database Systems*, vol. 4, no. 2, pp. 103-120, 2010.

[25] W.F. Battig, and W.E. Montague, "Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms," *Journal of Experimental Psychology Monographs*, vol. 80, PP. 1–46, 1969.

[26] J.P. Van Overschelde, K.A. Rawson, and J. Dunlosky (2004), "Category norms: An updated and expanded version of the Battig and Montague (1969) norms," *Journal of Memory and Language*, vol. 50, pp. 289–335, 2004.

[27] B. Munir, *AL-MAWRID: A Modern English-Arabic Dictionary*. Dar EL-ILMILMALAYIN, Beirut, Lebanon. Edition 11, 1977. www.malayin.com.

[28] J. Sinclair, *Collins Cobuild English Dictionary for Advanced Learners*, 3rd edn. Harper Collins, New York, 2001.

[29] S. Elkateb, W. Black, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum, "Building a WordNet for Arabic," *In Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

International Science Index, Cognitive and Language Sciences Vol:6, No:10, 2012 waset.org/Publication/12864