A New Feature-Based Method for Similarity Measurement under the Linux Operating System

https://doi.org/10.3991/ijim.v16i18.34455

Faaza A. Almarsoomi(✉), Israa A. Alwan
Department of Computer Science, University of Baghdad, Baghdad, Iraq
faaza.a.a@ihcoedu.uobaghdad.edu.iq

Abstract—This paper presents a new algorithm in an important research field which is the semantic word similarity estimation. A new feature-based algorithm is proposed for measuring the word semantic similarity for the Arabic language. It is a highly systematic language where its words exhibit elegant and rigorous logic. The score of sematic similarity between two Arabic words is calculated as a function of their common and total taxonomical features. An Arabic knowledge source is employed for extracting the taxonomical features as a set of all concepts that subsumed the concepts containing the compared words. The previously developed Arabic word benchmark datasets are used for optimizing and evaluating the proposed algorithm. In this paper, the performance of the new feature-based algorithm is compared against the performance of seven ontology-based algorithms adapted to Arabic. The results of the evaluation and comparison experiments show that the new proposed algorithm outperforms the adapted word similarity algorithms on the Arabic word benchmark dataset. The proposed algorithm will be included in the AWN-similarity which is free open-source software for Arabic.

Keywords—semantic similarity, feature-based algorithm, Arabic word similarity, and Arabic wordnet

1 Introduction

There is a massive increase in textual data generated over time resulting in increased interest in natural language processing applications particularly from Artificial Intelligence experts to enable the retrieval and analysis of this huge amount of data. Semantic word similarity estimation is an important component that plays a crucial role in text processing and understanding. It has been included in a wide range of natural language processing tasks intending to make them seem more intelligent [1] such as ontology learning, information retrieval, question answering, word sense disambiguation, text classification ,text summarization, electronic learning, and machine translation [2, 3, 4, 5]. The computational methods of semantic similarity are used for identifying and quantifying likeness between pair of words by exploiting the common attributes shared between them. These methods rely on the evaluation of the semantic information extracted from a knowledge source. Ontologies have been re-

ceived great attention from the research community concerned with semantic similarity. Several ontology-based methods have been proposed which can be categorized into methods rely on the structure of an ontology, methods that combine information content obtained from a corpus with semantic information, and methods rely on sets of ontological features. When it comes to semantic similarity, studying the features of a word considers very important due to the valuable information related to knowledge about the word that covered by this study. Features-based methods utilize semantic information more than those were exploited by methods in structure-based category. The additional knowledge contributes to better differentiate pairs of words and thus improves the similarity score [6]. Moreover, feature-based methods do not require sense-tagged data as those used by methods in the information content-based category. Where the process of producing this data is performed manually resulting in hindering the scalability and applicability of the information content-based methods with large corpora [6, 7]. In this paper, a new feature-based algorithm is proposed for measurement the word similarity for Arabic. Arabic is one of the Semitic languages which is the official language of the Arabic nation. The structure of the Arabic words is very systematic as it exhibits elegant and rigorous logic where the words are produced on the basis of a system of roots and templates [8]. For the Arabic language, the word similarity estimation field is more challenging because of the language's subtlety and higher complexity resulting in considering it a language with low semantic resources [9]. This explains the reason for the paucity of works presented in this field. The new proposed algorithm estimates the semantic similarity by taking use of taxonomical features omitted by other ontology-based algorithms. It utilizes a knowledge source for Arabic called as Arabic wordnet [10] as the base for performing the similarity estimation. The performance of the new feature-based algorithm is evaluated and compared against seven ontology-based algorithms using an Arabic word benchmark dataset. These ontology-based algorithms were previously introduced for English and then adapted to Arabic by [11] to develop open source packages to estimate Arabic word similarity that are free to use. The proposed algorithm aims to rival the adapted ontology-based algorithms in terms of accuracy.

Finally, the proposed algorithm will be included in the AWN-similarity packages [11] which are free open sources software that implement a number of word/concept similarity measures adapted to Arabic. All the implemented measures are on the basis of the structure and content of the AWN.

This paper is organized to review the existing ontology-based similarity measurements in section 2, followed by section 3 which introduces the new Arabic feature-based similarity algorithm. Section 4 illustrates the evaluation procedure of the proposed algorithm with a discussion of the experimental findings.

2 Ontology-based similarity measurements

A comprehensive discussion of semantic similarity techniques can be found in [1]. The existing similarity metrics created based on the English WordNet content and structure will be reviewed based on the following classification.

2.1 Structure-based similarity measures

Structure-based and also known as edge counting-based similarity measures typically use a function for computing the similarity based on the structure of ontology such as wordnet. The simplest form of the structure-based measures proposed by Rada et al. [12] where the similarity score was identified by calculating the distance (path length) between two concepts. The concepts with minimum path length considers more similar because they are close to each other in the taxonomy.

Leacock and Chodorow [13] presented a measure that estimating the similarity score by calculating the minimum path linking the concepts to be compared as well as the maximum depth in the taxonomy of English wordnet.

Wu and Palmer [14] proposed another structure-based measure for purpose of translating verbs from English to Chinese. This measure calculated the minimum path length linking the compared concepts by determining their meeting point that subsume them and the depth from meeting point to the root of taxonomy.

The similarity score identified by Pekar and Staab [15] measure was on the basis of the calculation of the minimum path between the compared concepts. The semantic similarity was directly proportional to the count of edges between the root of the taxonomy and the meeting point of the compared concepts.

Zhong et al. [16] used the depth from meeting point to the root of taxonomy with the depth of the compared concepts to identify the similarity score. A value denoted as milestone was given to each concepts in the hierarchy. This value was used to calculate the distance between each of the compared concept and their meeting point.

2.2 Information content-based similarity measures

The measures proposed for this group augmented the concepts of ontology with information content derived from a corpus. For each concept, its information content was calculated based on its occurrence in a corpus.

Resnik 1995 [17] proposed a similarity measure that computes the information content value of the concept that subsumes the concepts to be compared to determine the score of similarity. This measure gives the same similarity score to the concepts that have the same meeting point. This limitation was addressed by Jiang and Conrath [18] and Lin [19]. In their proposed measures some modifications were made for considering the information content of the concepts to be compared.

Lin's method compared the information content of the concept that subsumes the concepts to be compared as Resink with the information content values calculated for each concept.

Jiang and Conrath method calculated the path length linking the concepts to be compared linearly. For each concept, its information content was calculated and their sum was subtracted from the information content of the concept that subsumes the concepts to be compared.

2.3 Feature-based similarity measures

The proposed measures in this group determine the similarity score as a function of the evaluated concepts properties. These concepts are described as set of concepts that indicates their ontological features such as synonyms sets, glosses in wordnet and different semantic relationships. The similarity is determined based on the degree of overlap between the ontological features sets associated with each of the compared concepts.

Tversky 1977 [20] proposed a feature-based measure calculating the semantic similarity using the common and distinctive features of the two concepts. This measure does not consider the concept's position in taxonomy or its information content. The compared concepts were described as sets of words that indicate their features. Based on the notion that common features cause the similarity to increase while distinctive features tend to reduce it, the similarity score was calculated using Eq. (1).

$$Sim\ (a,b) = \propto \cdot F(A \cap B) - \beta \cdot F(A - B) - \gamma \cdot F(B - A) \quad (1)$$

Where a and b are the compared concepts, A and B are the sets of the compared concepts features. α, β and Υ represent the contribution of each of the proposed measure's component.

Rodriguez and Egenhofer [21] presented a feature-based measure that can be applied to estimate single or cross ontology similarity. The similarity was determined by combining the similarities between synonym sets, the distinguishing features, and semantic neighborhoods of the compared terms. The semantic similarity score was identified using Eq. (2).

$$Sim\ (a,b) = w \cdot\ S_{synsets}(a,b) + u \cdot\ S_{features}(a,b) + v \cdot\ S_{neighborhoods}(a,b) \quad (2)$$

Where w, u and v are the weight factors which represent the contribution of each of the proposed measure component. S is the overlap between different features used by this measure which calculated as

$$S\ (a,b) = \frac{|\ A \cap B|}{|A \cap B| + \gamma(a,b)|A \setminus B| + (1-\gamma(a,b))|B \setminus A|} \quad (3)$$

Where A, B are the sets of terms for a and b respectively, A\ B represent the terms belong to A but not belong to B. γ(a, b) is calculated as a function of the depth of the compared terms in the taxonomy using following Equation.

$$\gamma(a,b) = \begin{cases} \frac{depth(a)}{depth(a)+depth(b)}, depth(a) \leq depth(b) \\ 1 - \frac{depth(a)}{depth(a)+depth(b)}, depth(a) > depth(b) \end{cases} \quad (4)$$

Petrkis et al. [22] presented a feature-based measure known as X-similarity which calculated the similarity based on the matching process between synonym sets and glosses of concepts derived from English wordnet. The two terms were considered similar if their synonym sets, glosses and their neighborhoods concepts are lexically similar. The semantic similarity was identified using Eq. (5).

$$Sim(a,b) = \begin{cases} 1, \ if \ S_{synset}(a,b) > 0 \\ max\begin{cases} S_{neighborhoods}(a,b), S_{glosses}(a,b)\}, \\ if \ S_{synsets}(a,b) = 0 \end{cases} \end{cases} \quad (5)$$

The semantic similarity of the synonym sets Ssynsets and the glosses Sglosses were calculated as:

$$S(a,b) = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

Where A and B are the set of glosses o synonym sets for a and b, respectively.

The similarity of the neighbor concepts $S_{neighborhood}$s was calculated as:

$$S(a,b) = \max \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (7)$$

Where different kind of semantic relations are calculated separately and then the maximum is selected.

A non-linear measure was proposed by Sánchez et al. [6] for determining the similarity based on the concepts of semantic distance. In their proposed measure, the concepts sharing many generalizations in common have a distance shorter than those that have a smaller amount. Where, the semantic similarity was determined by comparing the distinctive taxonomic subsumers of the concepts to be compared with the sum of the taxonomic subsumers of each concept.

$$dis(c,d) = log_2\left(1 + \frac{|C \setminus D| + |D \setminus C|}{|C \setminus D| + |D \setminus C| + |C \cap D|}\right) \quad (8)$$

Where C and D are the taxonomical features (subsumers) set of c and d.

Likavec et al. [7] converted the structured-based measure proposed by Li et al. [23] into feature-based measure known as Sigmoid similarity measure. The proposed algorithm was evaluated using two datasets in the domain of recipes and Drink ontology. The performance of the sigmoid measure was compared with 3 structure-based similarity measures using the two datasets. These are Wu and Palmer, Li and Leacock and Chodorow measures.

$$Sim(a,b) = \frac{e^{CF(a,b)} - 1}{(e^{CF(a,b)} + 1)(DF(a) + DF(b) + 1)} \quad (9)$$

Where, CF and DF represent the common features and distinctive features, respectively. All the reported measures are for English and very few were presented for Arabic. Almarsoomi et al. [24] presented an Arabic word similarity (AWSS) measure inspired by Li measure [23]. AWSS is a structure-based algorithm where the semantic similarity score estimated as a function of two factors. These are the shortest path linking the words to be compared and the depth of the concept that subsumed them. The length and depth were derived from an Arabic knowledge source AWN which is a lexical database for Arabic created based on the design and content of a lexical database for English [25]. Figure 1 shows a part of AWN hierarchy.

This work focuses on the feature-based semantic similarity methods. These methods consider both common and distinctive features, hence utilizing semantic information more than those were exploited by methods in structure-based category. The additional knowledge contributes to better differentiate pairs of words and thus improves the similarity score [6, 7]. Moreover, feature-based measures do not require sense-tagged data as those used by methods in the information content-based category. Where the process of producing this data is performed manually resulting in hindering the scalability and applicability of the information content-based methods with large corpora.

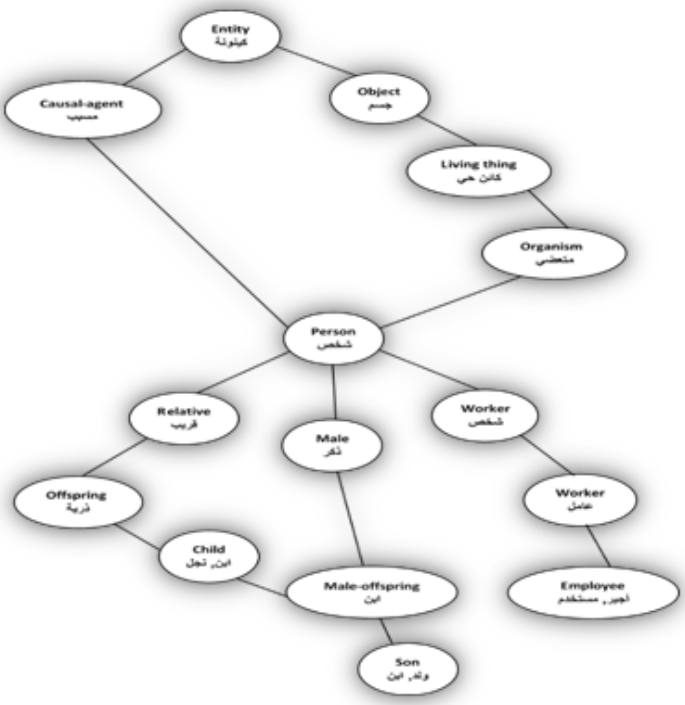

Fig. 1. A part of AWN hierarchy

3 The new feature-based algorithm

Feature-based methods estimate the similarity using semantic features that are hardly obtained in domain ontologies [6], such as concept properties (definitions, glosses, synsets) or non- taxonomic relationships. Accordingly, the possibility of applying these measures relied on the availability of this kind of information. An example of this issue, the glosses of Arabic senses are not available in the current version of AWN. Following [6], the decision was made to use taxonomical features extracted from AWN to estimate the score of semantic similarity. For each of the

compared concepts, a set of all its subsumers that indicates its taxonomical features will be associated. In accordance with the portion of hierarchy in Figure 1, the set of taxonomical features generated for the concept son ابن is:

(ابن, ابن, ذكر, نجل, ذرية, قريب, شخص, متعضي, كائن حي, جسم, مسبب, كينونة)

(Son, male-offspring, male, child, offspring, relative, person, organism, living things, object, casual-agent, entity)

While the set of features generated for the concept employee مستخدم is:

(مستخدم, عامل, شخص, متعضي, كائن حي, جسم, مسبب, كينونة)

(employee, worker, person, organism, living things, object, casual-agent, entity).

Psychological studies demonstrated that human estimate the similarity between words by comparing their similarities (common features) rather than their differences (distinctive features) [20]. Moreover, Tversky [20] illustrated that the common features of concepts cause the similarity to increase while distinctive features tend to reduce it. Taking these assumptions into consideration, in this research, the similarity of the concepts to be compared is determined as a function of their common and total features.

Given two Arabic concepts c1 and c2, their semantic similarity score is calculated as the ratio between the degree of overlap between the sets of taxonomical features associated with each concept and the total taxonomical features of the compared concepts.

$$Sim(c_1, c_2) = \frac{|X \cap Y|}{|X \cup Y|} \tag{10}$$

Where, X and Y represent the sets of concepts (all the concept's subsumers, including itself) that indicate the taxonomical features of c1 and c2, respectively. |X ∩ Y| represents the common features of the compared concepts and |X ∪ Y| represents their total taxonomical features. For our example, the common features between the concepts son ابن and employee مستخدم is 6 (شخص, متعضي, كائن حي, جسم, مسبب, كينونة) while their total features is 14 (ابن, ابن, ذكر, نجل, ذرية, قريب, شخص, متعضي, كائن حي, جسم, مسبب, كينونة, مستخدم, عامل)

It is intuitive that the concepts c1 and c2 should be more similar if they share more common features or they have fewer distinctive features (features belong to a concept but not belong to others). Consequently, the similarity should be directly proportional to the common features and inversely proportional to the distinctive features. To meet these requirements, the Equation of similarity in (10) is rewritten as follows:

$$\mathrm{Sim}(c_1, c_2) = \frac{\alpha\,|X \cap Y|}{\beta\,(|X \cup Y| - |X \cap Y|) + \alpha\,|X \cap Y|} \tag{11}$$

Where, α and β are the weight factors of the common and distinctive features. (|X ∪ Y| - |X ∩ Y|) represents the distinctive features of the compare concepts. The distinctive features of the concepts son ابن and employee مستخدم are 8 (ابن, ابن, ذكر, نجل, ذرية, قريب, مستخدم, عامل)

The similarity values of the proposed measure belong to (0, 1). If there are no common features between the compared concepts (|X ∩ Y| = 0), then the Similarity is equal to 0 ($Sim(c_1, c_2) = 0$) which fulfills the case of directly proportional. While in

the case of equivalent concepts, there are no distinctive features between the compared concepts and in this case the similarity is equal to 1, ($Sim(c_1, c_2) = 1$). This fulfills the case of inversely proportional. Most applications that used the similarity measures do not have sense tagged data. Consequently, the proposed algorithm should estimate the similarity score between Arabic words rather than Arabic concepts. Following Resnik [17], the similarity of the compared words is computed by the use of the pair of concepts for the two Arabic words resulting in maximum concept similarity.

$$\mathrm{Sim}(\mathrm{w}_1, \mathrm{w}_2) = \max\big(\mathrm{Sim}(\mathrm{c}_1, \mathrm{c}_2)\big), \quad \mathrm{c}_1 \epsilon\, \mathrm{s}(\mathrm{w}_1) \text{ and } \mathrm{c}_2 \epsilon\, \mathrm{s}(\mathrm{w}_2) \tag{12}$$

Where, $s(w_i)$ is the set of concepts in the AWN hierarchy which are senses of word w_i.

4 Evaluation procedure

4.1 Dataset

The validity of the new algorithm is not an easy task, assumed that the similarity measure's notion is a subjective human judgment. Resnick [26] demonstrated that the validity of an automated semantic similarity algorithm can be determined by comparing the performance of the algorithm with human cognition using a standard dataset. The first-word dataset with 70 Arabic word pairs was produced by [27] in 2012 and is known as ASSN-70. In its creation methodology, two experiments were conducted to generate the dataset word pairs and collect human ratings using a sample of 82 participants belonging to several Arabic countries. With the participation of 22 Arabic native speakers, 70 word pairs were produced using ordinary Arabic words selected from Arabic category norms. Human ratings for each of the produced Arabic word pair were collected by running a further experiment using a sample of 60 participants. Each pair of words was rated based on its similarity of meaning with a scale from 0 to 4. The similarity ratings for each of the produced word pairs were computed as the mean of the ratings given by 60 participants.

Two sub-datasets were created by partitioning the ANSS-70 for the purpose of using them in the process of optimizing and evaluating the new proposed methodologies [24]. These datasets are known as training and evaluation datasets where each one contains 35 pairs of Arabic words. In our experiment, the training dataset is employed for identifying the optimum parameters values of the new feature-based algorithm while the evaluation dataset is employed to assess its accuracy.

4.2 Normalization process

In this research, the latest version of AWN is utilized as the base for performing the similarity estimation. The words in AWN have been saved as lemmata with full diacritics while the modern words in Arabic are written without diacritics. Also, some letters in the Arabic language are written with marks such as (madda (~), and hamza

(ء)). The Arabic words with these letters were stored with marks in the AWN whilst they are written without marks in the modern writing system. These issues pose an interesting challenge to the automatic processing of semantic similarity algorithms for Arabic words that rely on the AWN hierarchy. To address these problems, a normalization process is performed to remove the diacritics and the marks from AWN words. Consequently, the Arabic words can be retrieved from AWN.

4.3 Optimum values of the weight factors

The weight factors of the new feature-based algorithm require identifying their optimal values. The training dataset described in section 4.1 was utilized to explore the role of the factors of the new algorithm. After assigning initial value to each weight factor, the proposed algorithm was used to generate machine rating for each word pair on the training dataset. The next step is to calculate the correlation coefficient between the ratings produced by the proposed algorithm with the mean of human judgments. A set of correlation coefficients was obtained by increasing the values of the weight factors α and β. Finally, the values that obtained the highest correlation were selected as the optimum values for the weight factors of the proposed algorithm. The optimum values achieved the strongest correlation are $\alpha = 0.5$ and $\beta = 0.8$.

4.4 AWN-similarity packages

The AWN-similarity packages are free open sources software which implemented a number of word/concept similarity measures adapted to Arabic. All the implemented measures are based on the AWN's structure and content. Seven measures were included into the newly produced packages. One of them was produced for Arabic (AWSS) and six measures were presented for English and successfully adapted to Arabic including Rada et al., Leacock and Chodorow (*Lch*), Wu Palmar (*Wup*), Pekar and Staab (*Pks*), Sánchez et al, and Zhong et al.

The developed packages were created to support and encourage Arabic researchers in the field to develop, validate and compare new methodologies with measures implemented in these packages. In our experiment, the performance of the new feature-based algorithm will be compared against the seven ontology-based algorithms on the evaluation benchmark dataset.

4.5 The proposed algorithm accuracy

Using the identified optimum values of the weight factors $\alpha = 0.5$ and $\beta = 0.8$, the accuracy of the new feature-based algorithm is evaluated by applying it to the evaluation dataset mentioned in section 4.1. The findings of the evaluation process are shown in Table 1 which represents the human ratings on the evaluation dataset and the ratings of the new feature-based measure.

In order to compare the performance of the new algorithm against the performances of the similarity algorithms implemented in AWN-Similarity package under the same condition, the correlation coefficient of each implemented measure stated in

authors' experiments are collected. The correlation coefficient of the new feature-based algorithm was computed between the machine ratings and the judgments provided by human on evaluation dataset. The findings of the evaluation and comparison experiments are illustrated in Table 2. Figure 2 shows the correlation coefficient achieved by the new algorithm and the implemented measures on evaluation dataset.

The boundaries of the possible performance expected from the computational methods of Arabic word semantic similarity have been calculated as the mean of the all Arabic participants' correlations (at 0.893) on the evaluation benchmark dataset as presented in Table 2.

The results in Table 2 show that the new feature-based algorithm performed very well which obtained a correlation coefficient at 0.90 that exceeded the average of the all participants' correlations at 0.893. The similarity ratings given by the new feature-based algorithm using the evaluation dataset are very close to those provided by humans on the same dataset as presented in Table 1.

Moreover, the proposed algorithm outperformed the similarity measures implemented in AWN-Similarity package which achieved the best correlation coefficient among them. Sanchez and AWSS algorithms achieved correlations close to the correlation obtained by new proposed algorithm as shown in Table 2.

Table 1. Similarity ratings on evaluation dataset from Human and New algorithm

No.	Word Pairs		Human Ratings	New Feature-based algorithm	أزواج الكلمـــات	
1	Coast	Endorsement	0.01	0	ساحـل	تصديـق
2	Noon	String	0.01	0.25	ظهـر	خيـط
3	Slave	Vegetable	0.04	0.05	عبد	خضـار
4	Smile	Village	0.05	0	ابتسامة/بسمة	قريــة
5	Hill	Pigeon	0.08	0.09	تــل	حمامة
6	Glass	Diamond	0.09	0.05	كأس	الماس
7	Cord	Mountain	0.13	0.17	حبـل	جبـل
8	Forest	Shore	0.21	0.12	غابــة	شاطئ
9	sepulcher	Sheikh	0.22	0.1	ضريـح	شيخ
10	Tool	Pillow	0.25	0.32	أداة	مخـدة
11	Coast	Mountain	0.27	0.45	ساحل	جبـل
12	Tool	Tumbler	0.33	0.56	أداة	قـدح
13	Journey	Shore	0.37	0	رحلة	شاطئ
14	Coach	Travel	0.40	0	حافلة	سفر
15	Feast	Fasting	0.49	0.17	عيـد	صيـام
16	Coach	Means	0.52	0.38	حافلة	وسيلـة
17	Girl	Sister	0.60	0.43	فتــاة	اخـت
18	Master	Sheikh	0.67	0.56	سيـد	شيخ

19	Food	Vegetable	0.69	0.56	خضار	طعـام
20	Slave	Odalisque	0.71	1	جاريـة	عبـد
21	Run	Walk	0.75	0.61	مشي	جـري
22	Cord	String	0.77	0.76	خيط	حبـل
23	Forest	Woodland	0.79	1	أحراش	غابـة
24	Cushion	Pillow	0.85	1	مخدة	مسند
25	Countryside	Village	0.85	1	قرية	ريف
26	Coast	Shore	0.89	1	شاطئ	ساحل
27	Tool	Means	0.92	1	وسيلة	أداة
28	Boy	Lad	0.93	1	فتى	صبي
29	Sepulcher	Grave	0.94	1	قبـر	ضريح
30	Glass	Tumbler	0.95	1	قـدح	كـأس

Table 2. The performance of the new algorithm against the AWN-Similarity package

Algorithms	**Evaluation Dataset**
Average of the correlation of all participants	0.893
Lch algorithm	0.839
Rada algorithm	0.851
Pks algorithm	0.886
Wup algorithm	0.84
Zhong algorithm	0.887
Sanchez algorithm	0.897
AWSS algorithm	0.894
New feature-based algorithm	0.90

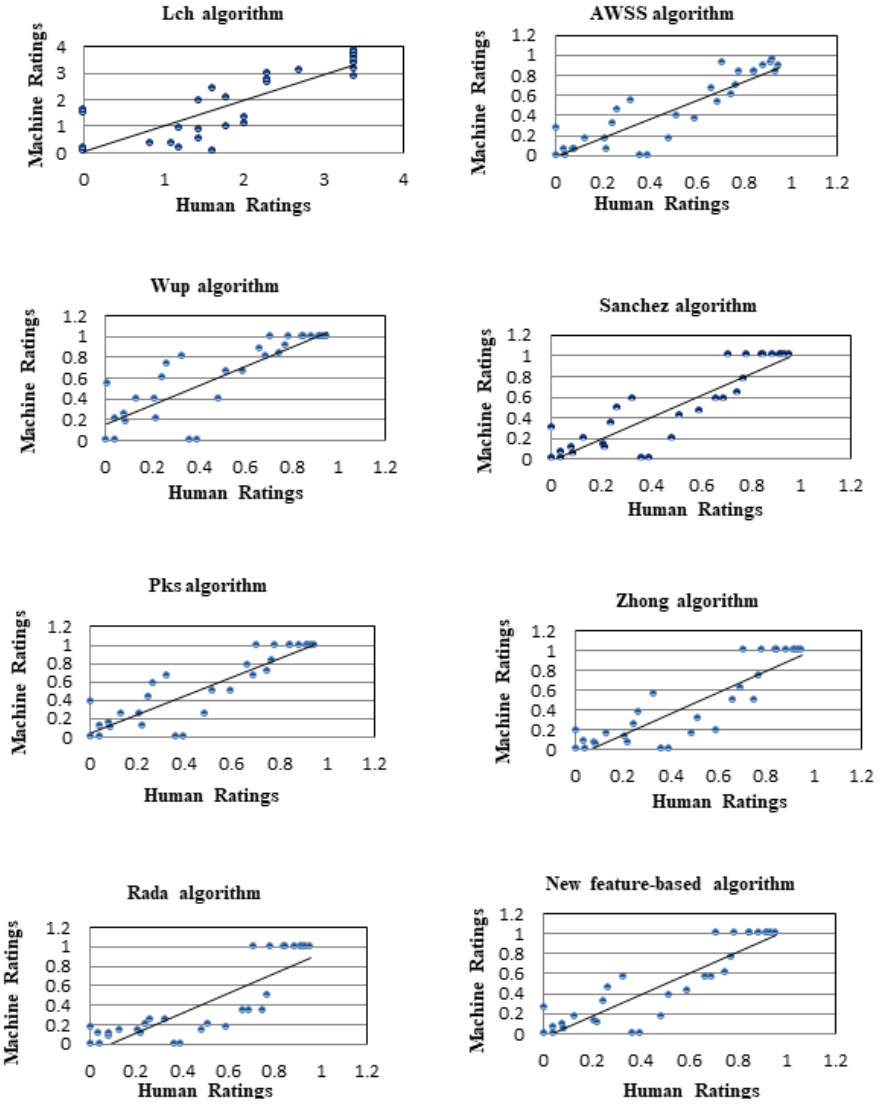

Fig. 2. The correlations achieved by the new feature-based algorithm and the implemented algorithms on evaluation dataset

5 Conclusion

This paper presented a computational word semantic similarity algorithm for a language with low semantic resources. Despite this issue, the proposed algorithm performed very well which gave similarity ratings close to those provided by humans resulting in achieving a good correlation coefficient at 0.90. The similarity was calculated by comparing the common features of the compared words with their total fea-

tures which emulate human estimation of word similarity. The features used by the proposed algorithm were derived from AWN as a set of taxonomical features for concepts that containing the compared words. Two experiments were performed to validate the new algorithm using two Arabic word similarity benchmarks datasets. The first experiment was conducted to identify the optimum parameters' values of the new algorithm using the training dataset while the second was for determining the algorithm accuracy on the evaluation dataset. The results of the validation experiment of the proposed algorithm suggest that a promising accuracy where it achieved a correlation better than the correlations reported by other algorithms implemented in AWN-Similarity. In the future, the feasibility of the proposed algorithm will be demonstrated by integrating the algorithm in real-live applications, in particular, the Arabic short answer grading system.

6 References

[1] D. Chandrasekaran, and V. Mago, "Evolution of Semantic Similarity - A Survey", *J. ACM*, vol. 37, no. 4, Article 111, 29 pages, 2020.

[2] T. Yuensuk, P. Limpinan, W. S. Nuankaew, and P. Nuankaew, "Information Systems for Cultural Tourism Management Using Text Analytics and Data Mining Techniques," *International Journal of Interactive Mobile Technologies,* vol. 66, no. 8, 2022. https://doi.org/10.3991/ijim.v16i09.30439

[3] A. Badawood and H. AlBadri, "Technology Based Model of a Mobile Knowledge as a Service to Facilitate Education Community," *International Journal of Interactive Mobile Technologies,* vol. 15, no. 24, 2021. https://doi.org/10.3991/ijim.v15i24.27335

[4] I. D. Wahyono, D. Saryono, H. Putranto, K. Asfani, H. A. Rosyid, M. M. Mohamad, M. N. H. B. M. Said, G. J. Horng, and J.-S. Shih, "Shared Nearest Neighbour in Text Mining for Classification Material in Online Learning Using Mobile Application," *International Journal of Interactive Mobile Technologies,* vol. 16, no. 4, 2022. https://doi.org/10.3991/ijim.v16i04.28991

[5] S. Mykytiuk, T. Moroz, S. Mykytiuk, M. Moroz, and O. Dolgusheva, "Seamless Learning Model with Enhanced Web-Quizzing in the Higher Education Setting," *International Journal of Interactive Mobile Technologies,* vol. 16, no. 03, p. 5, 2022. https://doi.org/10.3991/ijim.v16i03.27257

[6] D. Sánchez, M. Batet, D. Isern, and, A. Valls, "Ontology-based semantic similarity: A new feature-based approach", *Expert Systems with Applications*, vol 39, pp. 7718-7728, 2012. https://doi.org/10.1016/j.eswa.2012.01.082

[7] S. Likavec, I. Lombadi, and F. Cena, "Sigmoid Similarity- A New Feature-based Similarity Measure", *information science*, vol 481, pp. 203-218, 2019. https://doi.org/10.1016/j.ins.2018.12.018

[8] A. Farghaly, and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 14, 2009. https://doi.org/10.1145/1644879.1644881

[9] N. Y. Habash, "Introduction to Arabic Natural Language Processing," *Synthesis Lectures on Human Language Technologies*, vol. 3, pp.1-187, 2010. https://doi.org/10.2200/S00277ED1V01Y201008HLT010

[10] S. Elkateb, W. Black, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum, "Building a Wordnet for Arabic," In: *Proc. of the fifth international conference on Language Resources and Evaluation (LREC)*, 2006.
[11] F. A. Almarsoomi, and I. A. Alwan, "AWN-Similarity: Towards Developing Free Open Source Frameworks for Measuring Arabic Semantic Similarity under Windows / Linux Operating Systems", *Periodicals of Engineering and Natural Science*, vol. 9, no. 1, pp. 184-193, 2021. https://doi.org/10.21533/pen.v9i1.1791
[12] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and Application of a Metric on Semantic Nets", *IEEE Transactions on Systems, Man, and Cybernetics*, vol 19, no. 1, pp. 17-30, 1989. https://doi.org/10.1109/21.24528
[13] C. Leacock, and M. Chodorow, "Combining local context and WordNet similarity for word sense identification", In *WordNet: An electronic lexical database*, The MIT press, pp. 265-283, 1998.
[14] Z. Wu, and M. Palmer, "Verb Semantics and Lexical Selection", In: *Proc. of the 32nd Annual Meeting of the Associations for Computational Linguistics (ACL'94)*, pp. 133-138, 1994. https://doi.org/10.3115/981732.981751
[15] P. Viktor, and S. Steffen, "Taxonomy Learning: Factoring the Structure of a Taxonomy into a Semantic Classification Decision", In: *Proc. of the 19th International Conference on Computational Linguistics COLING '02, 2002*, vol. 1, pp. 1-7, 2002.
[16] J. Zhong, H. Zhu, J. Li, and Y. Yu, "Conceptual graph matching for semantic search", In: *Proc. of the 10th International Conference on Conceptual Structures*, pp. 92-106, 2006. https://doi.org/10.1007/3-540-45483-7_8
[17] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", In: *Proc. the 14th International Joint Conference on Artificial Intelligence*, vol. 1, p.448–453, 1995.
[18] J. Jiang, and D. Conrath, (1997), "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", In: *Proc. International Conference on Research in Computational Linguistics, ROCLING X*, pp. 19-33, 1997.
[19] D. Lin, "An Information-Theoretic Definition of Similarity", In *Proc. of Fifteenth International Conference on Machine Learning, ICML*, pp. 296-304, 1998.
[20] A. Tversky, "Features of Similarity", *Psycological Review*, vol. 84, pp. 327-352, 1977. https://doi.org/10.1037/0033-295X.84.4.327
[21] M. Rodríguez, and M. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, pp. 442–456, 2003. https://doi.org/10.1109/TKDE.2003.1185844
[22] E. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, "X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies", *Journal of Digital Information Management,* vol. *4*, PP. 233-237, 2006.
[23] Y. Li, Z. Bandar, and D. Mclean, "An Approach for Measuring Semantic Similarity between Words using Multiple Information Sources", *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, pp.871-882, 2003. https://doi.org/10.1109/TKDE.2003.1209005
[24] F. Almarsoomi, J. O'Shea, Z. Bandar, and K. Crockett, "AWSS: an Algorithm for Measuring Arabic Word Semantic Similarity", In*: Proc. IEEE international conference on systems, man, and cybernetics, SMC*, Manchester, United Kingdom, pp.504–509, 2013. https://doi.org/10.1109/SMC.2013.92
[25] G. A. Miller, "WordNet: A Lexical Database for English", *Communications of the ACM*, vol. 38, no. 1, pp.39–41, 1995. https://doi.org/10.1145/219717.219748

[26] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language", *Journal of Artificial Intelligence Research*, vol. 11, pp. 95-130, 1999. https://doi.org/10.1613/jair.514

[27] F. Almarsoomi, J. O'Shea, Z. Bandar, and K. Crockett, "Arabic Word Semantic Similarity", In: *Proc. of ICALLL (WASET)*, vol. 70, pp. 87–95, 2012.

7 Authors

Faaza A. Almarsoomi received her B. Sc. and M.Sc. in computing from the University of Technology, Iraq. The PhD degree in computing from Manchester metropolitan university, U. K., in 2015. She is currently an assistant professor with the University of Baghdad. Her research interests include semantics similarity for Arabic language, conversational agents and image processing. (email: faaza.a.a@ihcoedu.uobaghdad.edu.iq).

Israa A. Alwan received her B. Sc. in computing from the University of Baghdad, Iraq. The M.Sc. in computing from the University of Technology, Iraq. She is currently a Senior lecturer with the University of Baghdad. Her research interests include image processing, semantics similarity for Arabic language, and conversational agents.

Article submitted 2022-08-04. Resubmitted 2022-09-01. Final acceptance 2022-09-01. Final version published as submitted by the authors.