IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 5, October, 2023, pp. 1311-1322
Journal homepage: http://iieta.org/journals/ria

Assessing Semantic Similarity Measures and Proposing a WuP-Resnik Hybrid Metric for Enhanced Arabic Language Processing

Tahar Dilekh[1,2*], Mohamed Abderrahmen Boulahia[1], Saber Benharzallah[1,2]

[1] Computer Science Department, University of Batna 2, Fesdis, Batna 05078, Algeria
[2] LAMIE Laboratory, University of Batna 2, Fesdis, Batna 05078, Algeria

Corresponding Author Email: tahar.dilekh@univ-batna2.dz

https://doi.org/10.18280/ria.370524

Received: 7 May 2023
Revised: 20 September 2023
Accepted: 26 September 2023
Available online: 31 October 2023

Keywords:
semantic similarity measures, ontologies, WordNet (WN), Arabic WordNet (AWN), hybrid measures, WuP measures, Resnik measures

ABSTRACT

The accurate quantification of semantic similarity among Arabic words presents a significant challenge in Natural Language Processing (NLP). This is a critical aspect for a wide array of text-centric applications, including recommendation systems, plagiarism detection, and information retrieval. Enhanced performance in searches and classification is achieved by simplifying concepts within machine processing and unifying words with close meanings. This research investigates the complexities of measuring semantic similarity in Arabic, a language with distinct features such as the absence of short vowels in written text that renders distinguishing words without vowel diacritics challenging for computing systems. The effectiveness of various semantic similarity metrics is meticulously evaluated in this study, with a specific focus on their applicability to Arabic WordNet and English WordNet. The challenges associated with using Arabic WordNet for measuring word similarity are illuminated, and an innovative metric, integrating the Wu-Palmer and Resnik metrics, is proposed to enhance result accuracy. The primary accomplishment of this research resides in the identification of an optimal semantic similarity metric with a reduced error rate, thereby boosting the precision of results in NLP. This pivotal advancement paves the way for more accurate semantic assessments and improved performance across a broad spectrum of applications.

1. INTRODUCTION

Measuring semantic similarity between words lies at the heart of natural language understanding, holding profound implications for various applications in the field of Natural Language Processing (NLP). The pursuit of a precise semantic similarity measure with minimal error rates constitutes a fundamental quest in this domain. This endeavor is not merely an academic exercise; it carries substantial significance for practical applications and the advancement of research. In this introduction, we delve into the motivations behind this research and the unique challenges posed by the Arabic language in this context.

The importance of determining the most accurate semantic similarity measure cannot be overstated for several compelling reasons. Firstly, a measure that exhibits a low error rate provides results that align closely with human assessments, thereby enhancing the precision of NLP tasks. This heightened precision is particularly pivotal in applications where accurate semantic similarity measurements are imperative. For instance, a text classification application that is predicated on the semantic expansion approach. This approach is rooted in knowledge and involves a thorough investigation of various linguistic attributes, including the morphological, semantic, and syntactic relationships of query terms. The approach adopted here involves the replacement of query terms with words that possess similar contextual meanings, which leads to organize textual data into categories based on their similarities [1, 2].

Furthermore, lower error rates signify a reduction in the ambiguity surrounding semantic similarity measurements. A precise measure can effectively distinguish between words or concepts that are similar and those that are dissimilar. This reduction in ambiguity translates to fewer false positives and negatives, significantly impacting the reliability of NLP systems.

The significance extends to the broader realm of research. A semantic similarity measure with a low error rate can serve as a benchmark, facilitating consistent and comparable results across different NLP studies. Researchers can confidently build upon and refine their work, fostering advancements in the field.

In real-world applications such as plagiarism detection and question-answering systems, precision is paramount. Deploying a measure with a low error rate instills confidence in these applications, where accuracy directly translates into tangible benefits.

However, despite the universal need for accurate semantic similarity measures, developing reliable metrics remains a formidable challenge, particularly for languages like Arabic. The Arabic language introduces unique complexities into the realm of NLP. One such challenge is the absence of short vowels, known as "chakla," in written texts. These short vowels, which are not part of the alphabet, play a vital role in distinguishing between words with different meanings that are otherwise spelled identically. For example, the Arabic words 'عَلَمَ,' 'عِلْم,' 'عَلَم,' and 'عَلَّم' all share the same written form "علم" but convey distinct meanings.

In light of these challenges and the overarching need for precise semantic similarity measures, this research embarks on a comprehensive investigation. Our primary objective is to identify the most accurate semantic similarity measure while minimizing error rates. This investigation encompasses a thorough evaluation of various semantic similarity metrics applied to both Arabic WordNet (AWN) and English WordNet (WN). Importantly, we acknowledge and address the inherent disparities between these linguistic resources.

Additionally, this research introduces an innovative hybrid measure, drawing inspiration from the WuP and Resnik metrics. The aim is to potentially surpass the performance of existing metrics, especially within the specific dataset employed in this study. Our research endeavors to provide a deep analysis of diverse semantic similarity measures, elucidating their strengths and limitations. We underscore the critical importance of using the correct synset, emphasizing its role in generating relevant and realistic results.

The remainder of this paper is organized as follows: We commence by delving into the intricacies of calculating semantic similarity for Arabic words and contrasting the structures of WN and AWN. Subsequently, we provide an overview of prior research in the realm of semantic similarity measures based on WN/AWN. Our experimental case study follows, wherein we conduct a rigorous comparative evaluation of various semantic similarity measures, assessing their suitability for use in both WN and AWN. Finally, we conclude by discussing the implications of our research and charting avenues for future exploration.

In summation, this research paper contributes valuable insights into the domain of semantic similarity measures and their applicability in the realm of NLP. Our innovative hybrid measure holds the potential to significantly enhance the accuracy of various applications, including information retrieval and text classification, and presents exciting opportunities for further investigation in diverse languages and domains.

2. RELATED WORKS OF SEMANTIC SIMILARITY MEASURES BASED ON WN/AWN

The *ontology* is a formal representation of knowledge within a domain and the relationships between its concepts. Ontology is used in various fields such as the Semantic Web, Artificial Intelligence, and Biomedical Informatics, and is a useful tool for measuring semantic similarity between words. It provides a shared language for modeling a domain and offers important information that cannot be obtained from simple dictionaries. Ontology refers to the collection of concepts that are utilized to describe and represent a specific domain [3]. Gruber defines ontology as an explicit specification of a conceptualization [4], while computer science defines it as a formal representation of knowledge in a hierarchical way [5].

There are several ontologies available, and WN is one of them, a widely used lexical database for knowledge-based semantic similarity methods in computational linguistics and natural language processing. WN is primarily based on synonyms, with different synsets attributed to words with different meanings, and it organizes nouns, verbs, adverbs, and adjectives into semantic relations, represented as a hierarchical structure. WN provides four commonly used semantic relations for nouns: hyponym/hypernym, part meronym/part holonym, member meronym/member, and holonym. The most common relation is the hyponym/hypernym (is-a) relation, which accounts for close to 80% of the relations [6-9].

WN organizes concepts into hierarchy way Which shows the relations and the type of these relations between the different concepts, Figure 1.

AWN is a resource for Modern Standard Arabic that is designed based on Princeton WordNet (PWN) and Euro WordNet (EWN) [10]. AWN is mapped onto the Suggested Upper Merged Ontology (SUMO) [10], which is a formal ontology that contains about 1000 Ontology Concepts, 4000 Ontology Axioms and 750 Ontology Rules [11]. AWN contains four entity types: item, word, form, and link. There is a significant addition in the number of words between version 2.0 and version 2.0.1 of AWN, which may affect the results of similarity measurement [10].

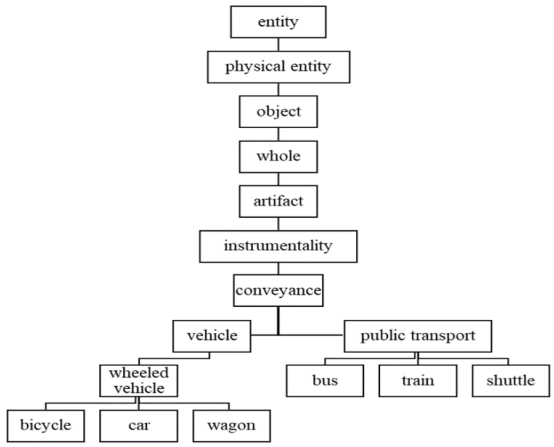

Figure 1. Sample hierarchical structure of fragment of WordNet

The challenge associated with the utilization of AWN primarily revolves around its limitations. AWN is relatively constrained, which is particularly notable given the vast lexical diversity within the Arabic language, surpassing that of many other languages. Additionally, the intricacies inherent in accurately pinpointing the appropriate synset for certain words compound the challenge. While methodologies exist for the identification of suitable synsets, the inherent complexity of distinguishing between closely related synsets remains a formidable obstacle for automated processes, often necessitating human intervention for resolution.

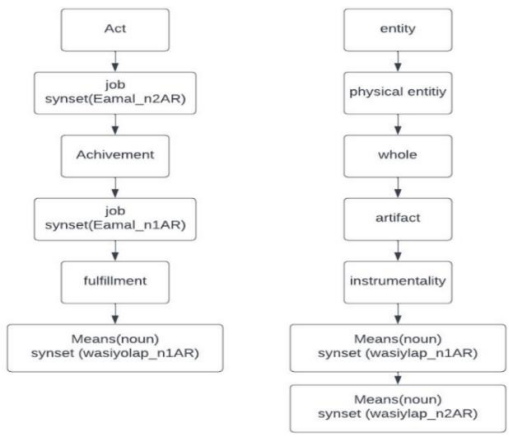

Figure 2. The different path between the synsets of the noun "Means"

Table 1. Summary of related works on semantic similarity measures based on WN/AWN

Category	Measures	Principle	Advantages	Disadvantages
Path based	The Shortest Path [12, 13], PATH measure [14]	Based only on the distance between two concepts	Simple to implement	The similarity of two pairs with equal distance between two concepts will be the same.
	Wu and Palmer [15], Almarsoomi et al. [16], Leakcock and Chodorow [17]	Based on the depths of concept and the length	Simple to implement	The similarity of two pairs with equal depths and length will be the same.
	Li et al. [18]	Based on the assumption that information sources are infinite to some extend	Simple to implement	
IC based [9, 19-25]	Resnik [21]	Based on the information content of lso	Simple to implement	The similarity of two pairs with the same lso will be the same similarity
	Jiang and Conrath [26], Lin [27]	Based on the information content of concepts and their lso	Consider the IC of the concepts that are being compared	The similarity of two pairs with the same sum of IC(c1) and IC(c2) will be the same.
	Lord et al. [28]	Based on the IC value given by Resnik	simple to implement	The similarity of two pairs with the same lso will be the same similarity
	Seco et al. [29]	Apply a linear transformation to the jiang &Conrath formula	Consider the IC of the concepts that are being compared	two pairs with the same information content that uses hyponymy will have the same similarity
	Saruladha et al. [22]	Consider hyponymy and meronomy of concepts	Provides an independent solution to the sparse data problem that is prevalent in corpus.	two pairs with the same information content that uses hyponymy and meronomy will have the same similarity
	Seddiqui and Aono [24]	considers the relation of properties, property function, and restrictions	Consider the relation of properties	The similarity of two pairs with the same sum of IC(c1) and IC(c2) and the same lso will be the same.
	Meng and Gu [30]	based on Lin's method	consider the IC of the concepts that are being compared	The similarity of two pairs with the same sum of IC(c1) and IC(c2) will be the same.
Feature-based [9, 31]	Tversky [32], Ezzikouri et al. [33]	Features shared by a subclass and its superclass contribute more to the similarity evaluation than features shared in the opposite direction.	Consider the features of the concept.	Issue of missing glosses in most of ontologies and problem of computational complexity.
	Lesk [34]	The overlap between the corresponding definitions of two concepts, be used to measure the relatedness between two concepts	Can be used in conjunction with any dictionary	Issue of missing glosses in most of ontologies and problem of computational complexity.
	Patwardhan and Pedersen [35], Patwardhan [36]	Uses context vectors to combine the glosses content of taxonomic concepts	Combine the glosses content of taxonomic concepts with statistical data extracted from the corpus	
	Merhbene et al. [37]	Modified the Lesk algorithm by using different semantic similarity measures.	Resolve the issue of missing glosses	
	Jiang et al. [38]	The similarity value is increased by common features, while the similarity value is decreased by non-common features	Determining semantic similarity based on the glosses of Wikipedia concepts.	
	Ezzikouri et al. [33]	The similarity value is increased by common features	Consider the features of the concept	
Hybrid method	Zhou et al. [39]	Based on lengths between and IC of concept	Based on various information from different categories	Depends on the categories that are combined.
	Aldiery [19]	Combine multiple information sources	Based on various information from different categories	

The problem of using Arabic WordNet is that we can't tell the difference between the synset 'wasiylap_n2AR' and the synset 'wasiylap_n1AR' because they are in the same path and the synset 'wasiylap_n2AR' is straight down the synset 'wasiylap_n1AR' which means if we choose the synset 'wasiylap_n2AR' we will get the depth increased by 1 and the distance between other synset also increased by 1 leading us to different result by the measure of similarity, as shown in

Figure 2. Which made it necessary to use WN instead of AWN in many cases.

WN is regarded as a valuable resource for identifying semantic similarity between two words due to its organization of words according to lexical relationships. Numerous similarity measures utilizing WN have been suggested.

Generally, the conventional similarity measures fall into four categories, which include path-based measures, information content-based measures, feature-based measures, and hybrid measures.

In Table 1, we aimed to summarize the most relevant works on semantic similarity measures based on WN/AWN, highlighting the advantages and limitations of each method.

3. EXPERIMENTAL CASE STUDY

The purpose of this section is to conduct a comparative evaluation of different measures of semantic similarity and assess their suitability for use in both WN and AWN. In this comparison, we will consider the differences between WN and AWN. The chosen measures will then be applied to an Arabic dataset and their results will be analyzed. The aim is to identify the measure that exhibits the best performance for automated tasks.

3.1 Arabic dataset benchmark

The dataset has been used in this study is the Arabic dataset benchmark created by Faza et al. [40]. The dataset was created following the same methods as the English dataset benchmarks for semantic similarity. The dataset was created in two stages. In the first stage, the Arabic word pairs set was selected and translated into Arabic using an English-Arabic dictionary. Two sets of Arabic noun pairs ranging from high similarity of meaning (HSM) to medium similarity of meaning (MSM) and low similarity were generated. In the second stage, the human similarity rate for word pairs was specified. 60 participants from various Arabic countries were asked to rank the set of 70 Arabic word pairs gathered in the first stage in order of importance. They ranked each word pair on a five-point scale ranging from 0.0 (unrelated) to 4.0 (the same). The dataset is important for evaluating semantic similarity between Arabic words.

Table 2 presents a sample of the data benchmark used in our study.

Table 2. A snippet of the benchmark dataset

N	English Word Pairs		Human Ratings	Arabic Word Pairs	
01	Coast	Endorsement	0.03	تصديق	ساحل
02	Noon	String	0.03	خيط	ظهر
03	Cushion	Diamond	0.06	الماس	مسند
04	Gem	Pillow	0.07	مخدة	جوهرة
05	Stove	Walk	0.07	مشي	موقد
06	Cord	Midday	0.08	ظهيرة	حبل
07	Signature	String	0.08	خيط	توقيع
08	Boy	Endorsement	0.12	تصديق	صبي
09	Boy	Midday	0.16	ظهيرة	صبي
10	Slave	Vegetable	0.16	خضار	عبد
11	Smile	Village	0.18	قرية	ابتسامة
12	Smile	Pigeon	0.20	حمامة	ابتسامة
13	Wizard	Infirmary	0.22	مشفى	ساحر
14	Noon	Fasting	0.29	صيام	ظهر

The column labeled "N" is represented by the following abbreviation: The numbering of the word.

3.2 Applicable measures on AWN

As mentioned earlier, most semantic similarity measures, including feature-based measures and corpus-dependent information content measures cannot be used on AWN due to its limitations. Table 3 illustrates the reasons why certain common semantic similarity measures are not applicable to AWN.

Table 3. Assessing the suitability of traditional semantic measures on AWN

Category	Measures	Applicable on AWN	Justifications
Path based	All Path Based Measures	Yes	AWN provides a path information source.
	Li	Yes	Non-linear function of the shortest path and depth of lso which are available in AWN
IC based	All current IC based Measure except Meng	No	Finding Arabic word dependent frequency with diacritics is difficult since data is few.
	M	Yes	Depth of LCS and max depth which are available in AWN
Feature-based	MF	Limited	Issue of missing glosses in most of ontologies and problem of computational complexity.
	Zo	Yes	It uses different semantic similarity measures to replace Lesk's original measure to find the gloss that corresponds to the correct sense of the ambiguous word.
Hybrid method	Zh, GA	Yes	Consists of a combination of applicable measures on AWN

"Measure" is represented by the following abbreviation: SP: The Shortest Path, PM: PATH measure, WP: Wu & Palmer's, F: Faaza, LC: Leakcock &Chodorow, Li: Li, R: Resnik, J: Jiang, L: Lin, M: Meng, MF: Most of feature-based measures, Zo: Zouaghi, Zh: Zhou and GA: Ghandi Aldiery.

3.3 Choosing synsets in AWN

We discussed the difficulty of selecting synsets in AWN, which led us to use two methods for selecting synsets: automatic selection by the program and manual selection. To ensure accurate results, we primarily chose synsets automatically, but some were manually corrected. In all comparisons, the same synset was chosen for each term to eliminate any variations that may arise from using different synsets. We specifically selected the synset of a concept instead of its synonyms. It is possible that our results may differ from those of other studies because we prioritized the

selection of specific synsets over the general ones.

In AWN's file, each synset has a unique name, but some synsets may have multiple values, making it difficult to locate the correct synset for a given concept. For example, in the Figure 3, if we search for the word "سفر", we may find multiple synsets, and while one may be correct, the other may provide a better similarity score. In our study, we present two results: one represents the best possible outcome, while the other represents the most realistic outcome for automated work, given the lack of a reliable method for determining the best synset.

Table 4. The selected synset for the first experiments (using the best possible synsets)

N	Word 1	Word 2	Synset (w1)	Synset (w2)	Human Rating	Depth (lso)	Depth (w1)	Depth (w2)	Len (w1, w2)
01	تَصْدِيق	سَاجِل	taSodiyq_n2AR	$aATi}_AlbaHor_n1AR	0.01	0	4	1	0
02	خَيْط	ظُهْر	xaATa_v1AR	<irotafaEa_v3AR	0.01	0	4	3	0
03	مَشْي	موقد	ma$oy_n1AR	Non	0.01				
04	خُضَار	عَبْد	xuDar_n1AR	xaAdim_n1AR	0.04	1	7	5	10
05	قَرْيَة	بَسْمَة	qaroyap_n1AR	basomap_n1AR	0.05	0	5	8	0
06	مشفى	ساحر	Non	Non	0.06				
07	حَمَامَة	تَلّ	HamaAm_n1AR	rukaAm_n1AR	0.08	2	11	6	13
08	الْماس	كَأْس	AlomAs_n1AR	kuwb_n1AR	0.09	1	7	7	12
09	جَبَل	حَبْل	jabal_n1AR	laq~aHa_v1AR	0.13	0	1	3	0
10	شَاطِئ	غابَة	$aATi}_AlbaHor_n1AR	gaAbap_n1AR	0.21	0	1	4	0
11	شَيْخ	ضَرِيح	qabor_n1AR	ra}iyos_n1AR	0.22	1	5	5	8
12	مِخَدَّة	أَدَاة	wisaAdap_n1AR	>aadaAp_n1AR	0.25	4	6	7	5
13	جَبَل	سَاجِل	jabal_n1AR	$aATi}_AlbaHor_n1AR	0.27	0	1	1	0
14	قَدَح	أَدَاة	kuwb_n1AR	>daAp_n1AR	0.33	5	7	6	3
15	شَاطِئ	رِحْلَة	$aATi}_AlbaHor_n1AR	Harakap_n1AR	0.37	0	1	4	0
16	سَفَر	حَافِلَة	taHar~uk_n1AR	HaAfilap_n1AR	0.4	0	5	8	0
17	فرن	طعام	Non	TaEaAm_n3AR	0.44				
18	صِيَام	عِيد	Sawom_n1AR	<iHotifaAl_n1AR	0.49	2	6	4	6
19	وَسِيلَة	حَافِلَة	wasiylap_n1AR	HaAfilap_n1AR	0.52	5	6	8	4
20	أُخْت	فَتَاة	>ax_n1AR	fataAp_n1AR	0.6	3	5	6	5
21	جبل	تَلّ	jabal_n1AR	rukaAm_n1AR	0.65	0	1	6	0
22	شَيْخ	سَيِّد	ra}iyos_n1AR	say~id_n1AR	0.67	3	5	6	5
23	خُضَار	طَعام	xuDaAr_n1AR	TaEaAm_n3AR	0.69	4	6	4	2
24	جَارِيَة	عَبْد	xaAdim_n1AR	Eabod_n1AR	0.71	3	5	4	3
25	مَشْي	جَرْي	ma$oy_n1AR	jaroy_n1AR	0.75	5	6	6	2
26	خَيْط	حَبْل	gazol_n1AR	Habol_n1AR	0.77	6	7	6	1
27	أَخْرَاش	غابَة	dagol_n1AR	dagol_n1AR	0.79	5	5	5	0
28	مِخَدَّة	مِسْنَد	wisaAdap_n1AR	wisaAdap_n1AR	0.85	6	6	6	0
29	قَرْيَة	رِيف	riyf_n1AR	riyf_n1AR	0.85	5	5	5	0
30	شَاطِئ	سَاجِل	$aATi}_AlbaHor_n1AR	$aATi}_AlbaHor_n1AR	0.89	0	1	1	0
31	وَسِيلَة	أَدَاة	wasiyolap_n1AR	>adaAp_n1AR	0.92	6	6	7	1
32	قَنَى	صَنَوِيّ	muraAhiq_n1AR	muraAhiq_n1AR	0.93	5	5	5	0
33	قَبْر	ضَرِيح	qabor_n1AR	qabor_n1AR	0.94	5	5	5	0
34	مشعوذ	ساحر	Non	Non	0.94				
35	قَدَح	كَأْس	kuwb_n1AR	kuwb_n1AR	0.95	7	7	7	0

```
<item itemid="safar_n1AR" offset="100282711"
name="سَفَر" type="synset" POS="n" source="AWNcore"
authorshipid="8361"/>
<word wordid="safar_1" value="سَفَر"
synsetid="safar_n1AR" frequency="0" corpus=""
authorshipid="30020"/>
<word wordid="tiroHaAl_1" value="تِرْحَال"
synsetid="safar_n1AR" frequency="1"
authorshipid="32763"/>
<item itemid="taHar~uk_n1AR" offset="100270493"
name="تَحَرُّك" type="synset" POS="n" source="AWNcore"
authorshipid="8875"/>
<word wordid="safar_3" value="سَفَر"
synsetid="taHar~uk_n1AR" frequency="" corpus=""
authorshipid="30022"/>
<word wordid="taHar~uk_1" value="تَحَرُّك"
synsetid="taHar~uk_n1AR" frequency="0" corpus=""
authorshipid="31157"/>
```

Figure 3. A snippet of AWN's xml file

3.4 Using conventional measures with AWN and WN

The aim of these experiments is to compare several widely used measures that can be applied to both AWN and WN. The purpose is to identify the WN and measure combination that produces superior outcomes.

3.4.1 Experiment 1: Comparison of measures on AWN with synset selection for optimal results

Selecting synset. In this comparison, we only considered synsets that yield superior outcomes, regardless of whether they are the correct synsets or not.

Table 4 reveals that numerous synsets are missing in AWN, such as "مشفى" ,"مشعوذ", and "موقد". Moreover, several pairs yield a similarity score of 1, indicating that both terms belong to the same synset, and some terms have no synset at all. However, as mentioned earlier, selecting synsets in this manner may lead to many errors. For instance, in pair number 4, the synset "xaAdim_n1AR" was selected for the concept "عبد", but in pair 24, the synset "Eabod_n1AR" was selected for the same concept. Although this approach may improve results, it lacks a synset selection methodology. In this experiment, we chose the synset that results in a smaller error based on the provided data.

Result. We evaluated six measures, as shown Table 5.

Table 5. Result of the first experiment (applying similarity measures on AWN using the best possible synsets)

N	HR	SP	W	Li	F	A	LC
01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
03	0.01	--	--	--	--	--	--
04	0.04	0.67	0.17	0.07	0.05	0.20	0.78
05	0.05	0.00	0.00	0.00	0.00	0.00	0.00
06	0.06	--	--	--	--	--	--
07	0.08	0.57	0.24	0.06	0.05	0.34	0.66
08	0.09	0.60	0.14	0.05	0.03	0.17	0.70
09	0.13	0.00	0.00	0.00	0.00	0.00	0.00
10	0.21	0.00	0.00	0.00	0.00	0.00	0.00
11	0.22	0.73	0.20	0.11	0.06	0.24	0.88
12	0.25	0.83	0.62	0.36	0.33	0.71	1.08
13	0.27	0.00	0.00	0.00	0.00	0.00	0.00
14	0.33	0.90	0.77	0.55	0.51	0.81	1.30
15	0.37	0.00	0.00	0.00	0.00	0.00	0.00
16	0.4	0.00	0.00	0.00	0.00	0.00	0.00
17	0.44	--	--	--	--	--	--
18	0.49	0.80	0.40	0.25	0.17	0.50	1.00
19	0.52	0.87	0.71	0.45	0.43	0.78	1.18
20	0.6	0.83	0.55	0.35	0.27	0.65	1.08
21	0.65	0.00	0.00	0.00	0.00	0.00	0.00
22	0.67	0.83	0.55	0.35	0.27	0.65	1.08
23	0.69	0.93	0.80	0.66	0.53	0.83	1.48
24	0.71	0.90	0.67	0.52	0.37	0.73	1.30
25	0.75	0.93	0.83	0.67	0.60	0.85	1.48
26	0.77	0.97	0.92	0.82	0.75	0.92	1.78
27	0.79	1.00	1.00	1.00	1.00	1.00	1.49
28	0.85	1.00	1.00	1.00	1.00	1.00	1.49
29	0.85	1.00	1.00	1.00	1.00	1.00	1.49
30	0.89	1.00	1.00	1.00	1.00	1.00	1.49
31	0.92	0.97	0.92	0.82	0.75	0.92	1.78
32	0.93	1.00	1.00	1.00	1.00	1.00	1.49
33	0.94	1.00	1.00	1.00	1.00	1.00	1.49
34	0.94	--	--	--	--	--	--
35	0.95	1.00	1.00	1.00	1.00	1.00	1.49

The first row is represented by the following abbreviations: N: the numbering of the word, HN: Human rating, SP: Shortest path, W: WuP measure, Li: Li measure, F: faza, A: Aldieryand LC: Leakcock & Chodorow.

The Table 6 shows the correlation and mean squared error for each measure in this comparison.

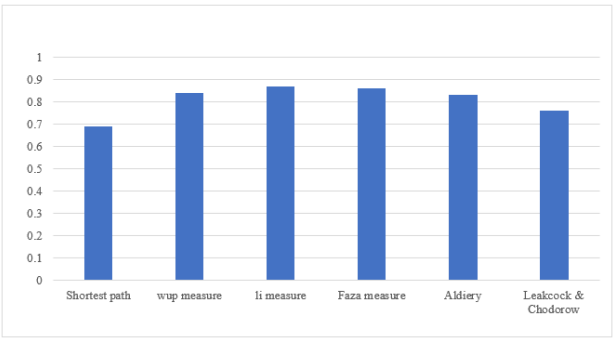

Figure 4. The correlation between human ratings and similarity measures scores in AWN experiment using the best possible synsets

Among the six measures utilized, the li measure had the highest correlation value, as shown in Figure 4, and the lowest mean squared error, indicating that it generated better similarity scores. It is unsurprising that the shortest path had the lowest correlation because it only takes into account the distance between concepts.

Table 6. The correlation and MSE of the measures in the first experiment

Measures	Correlation	MSE
Shortest path	0.69	0.104
Wup measure	0.84	0.046
Li measure	0.87	0.042
Faza measure	0.86	0.051
Aldiery	0.83	0.052
Leakcock & Chodorow	0.76	0.086

3.4.2 Experiment 2: Comparison of measures on AWN with synset selection of correct synsets

Selecting synset. For this experiment, we only used the right synsets for each concept, as shown in Table 7, even if they do not produce the best results, so the correlation value decreased in this experiment.

Table 7. The selected synset for the second experiments (Right synsets)

N	HR	Synset (w1)	Synset (w2)
01	0.01	taSodiyq_n2AR	$aATi}_AlbaHor_n1AR
02	0.01	xaATa_v1AR	<irotafaEa_v3AR
03	0.01	ma$oy_n1AR	Non
04	0.04	xuDar_n1AR	Eabod_n1AR
05	0.05	qaroyap_n1AR	basomap_n1AR
06	0.06	Non	Non
07	0.08	HamaAm_n1AR	rukaAm_n1AR
08	0.09	AlomAs_n1AR	kuwb_n1AR
09	0.13	jabal_n1AR	laq~aHa_v1AR
10	0.21	$aATi}_AlbaHor_n1AR	gaAbap_n1AR
11	0.22	qabor_n1AR	ra}iyos_n1AR
12	0.25	wisaAdap_n1AR	>aadaAp_n1AR
13	0.27	jabal_n1AR	$aATi}_AlbaHor_n1AR
14	0.33	kuwb_n1AR	>daAp_n1AR
15	0.37	$aATi}_AlbaHor_n1AR	Harakap_n1AR
16	0.4	safar_n1AR	HaAfilap_n1AR
17	0.44	Non	TaEaAm_n3AR
18	0.49	Sawom_n1AR	<iHotifaAl_n1AR
19	0.52	wasiylap_n1AR	HaAfilap_n1AR
20	0.6	>ax_n1AR	fataAp_n1AR
21	0.65	jabal_n1AR	rukaAm_n1AR
22	0.67	ra}iyos_n1AR	say~id_n1AR
23	0.69	xuDaAr_n1AR	TaEaAm_n3AR
24	0.71	xaAdim_n1AR	Eabod_n1AR
25	0.75	ma$oy_n1AR	jaroy_n1AR
26	0.77	xayoT_n1AR	Habol_n1AR
27	0.79	dagol_n1AR	gaAbap_n1AR
28	0.85	wisaAdap_n1AR	wisaAdap_n1AR
29	0.85	qaroyap_n1AR	riyf_n1AR
30	0.89	$aATi}_AlbaHor_n1AR	$aATi}_AlbaHor_n1AR
31	0.92	wasiyolap_n1AR	>adaAp_n1AR
32	0.93	muraAhiq_n1AR	Sabiy~_n1AR
33	0.94	qabor_n1AR	qabor_n1AR
34	0.94	Non	Non
35	0.95	kuwb_n1AR	kuwb_n1AR

The column labeled "N" is the abbreviation of the numbering of the word, and "HR" is the abbreviation of Human Rating

Result. Table 8 indicates that numerous synsets, such as the term "قدح" which shares the same synset as "كأس", are still missing in AWN. Additionally, a noteworthy observation is

the significant contrast in the degree of similarity between "قرية" and "ريف"; in the previous experiment, their similarity score was 1, but in this experiment, it dropped to zero since their correct synsets do not belong to the same hierarchy in AWN.

Table 8. Result of the second experiment (applying similarity measures on AWN using the correct synsets)

N	HR	SP	W	Li	F	A	LC
01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
03	0.01	--	--	--	--	--	--
04	0.04	0.70	0.18	0.09	0.05	0.22	0.82
05	0.05	0.00	0.00	0.00	0.00	0.00	0.00
06	0.06	--	--	--	--	--	--
07	0.08	0.57	0.24	0.06	0.05	0.34	0.66
08	0.09	0.60	0.14	0.05	0.03	0.17	0.70
09	0.13	0.00	0.00	0.00	0.00	0.00	0.00
10	0.21	0.00	0.00	0.00	0.00	0.00	0.00
11	0.22	0.73	0.20	0.11	0.06	0.24	0.88
12	0.25	0.83	0.62	0.36	0.33	0.71	1.08
13	0.27	0.00	0.00	0.00	0.00	0.00	0.00
14	0.33	0.90	0.77	0.55	0.51	0.81	1.30
15	0.37	0.00	0.00	0.00	0.00	0.00	0.00
16	0.4	0.00	0.00	0.00	0.00	0.00	0.00
17	0.44	--	--	--	--	--	--
18	0.49	0.80	0.40	0.25	0.17	0.50	1.00
19	0.52	0.87	0.71	0.45	0.43	0.78	1.18
20	0.6	0.83	0.55	0.35	0.27	0.65	1.08
21	0.65	0.00	0.00	0.00	0.00	0.00	0.00
22	0.67	0.83	0.55	0.35	0.27	0.65	1.08
23	0.69	0.93	0.80	0.66	0.53	0.83	1.48
24	0.71	0.90	0.67	0.52	0.37	0.73	1.30
25	0.75	0.93	0.83	0.67	0.60	0.85	1.48
26	0.77	0.87	0.67	0.44	0.38	0.74	1.18
27	0.79	0.77	0.22	0.13	0.07	0.27	0.93
28	0.85	1.00	1.00	1.00	1.00	1.00	1.49
29	0.85	0.00	0.00	0.00	0.00	0.00	0.00
30	0.89	1.00	1.00	1.00	1.00	1.00	1.49
31	0.92	0.97	0.92	0.82	0.75	0.92	1.78
32	0.93	0.97	0.89	0.81	0.62	0.89	1.78
33	0.94	1.00	1.00	1.00	1.00	1.00	1.49
34	0.94	--	--	--	--	--	--
35	0.95	1.00	1.00	1.00	1.00	1.00	1.49

The first row is represented by the following abbreviations: N: the numbering of the word, HN: Human rating, SP: Shortest path, W: WuP measure, Li: Li measure, F: faza, A: Aldieryand LC: Leakcock & Chodorow.

We observe a more substantial discrepancy between the human ratings and the measurement outcomes in this experiment compared to the previous one, indicating an elevated MSE and a reduction in correlation. Table 9 demonstrates that the ranking of similarity measures is not significantly different from the previous experiment, as the li measure continues to yield the most favorable results in terms of both correlation and MSE.

We can observe that all similarity measures exhibit a decrease in correlation and an increase in MSE, highlighting the impact of synset selection on achieving accurate results. The Table 8 presents a more realistic outcome for automated similarity measures, providing a clearer picture of the extent of variation between each measure (Figure 5).

Table 9. The correlation and MSE of the measures in the second experiment

Measures	Correlation	MSE
Shortest path	0.58	0.126
Wup measure	0.72	0.077
Li measure	0.75	0.081
Faza measure	0.73	0.097
Aldiery	0.70	0.082
Leakcock & Chodorow	0.66	0.105

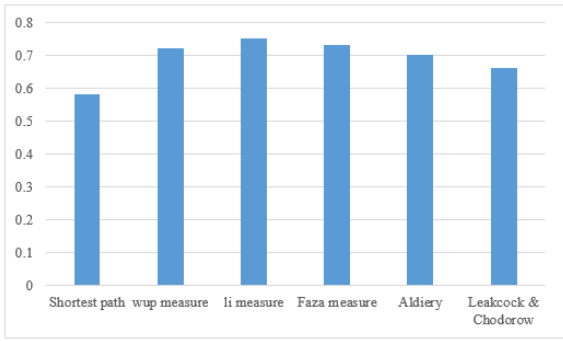

Figure 5. The correlation between human ratings and similarity measures scores in AWN experiment using the correct synsets

3.4.3 Experiment 3: Comparison of traditional measures on WN

Selecting synset. We utilized the NLP library that incorporates the English WN to pick the synset in the WN. It is noteworthy that in the English WN, finding the synset of a concept was more straightforward since it has a greater number of synsets than the AWN, and it provides a more precise definition of synsets because the term is accessible in both Arabic and English, thereby mitigating the issue of homonyms (Table 10).

Result. We compared five different similarity measures from various categories on the English WN, including Path, Wu & Palmer's, Leacock & Chodorow, Resnik, and Lin measures. Table 11 presents a comparison of the most widely used similarity measures in WN, with WuP, Path, and Leacock & Chodorow belonging to the path category and Resnik and Lin to the information content - corpus dependent category. This category was not applicable in AWN due to the unavailability of the required corpus in Arabic. Upon initial observation, we noticed that the English WN includes all the necessary concepts for this experiment, and each word has its synset, unlike AWN, where the issue of synset deficiency arose.

Table 10. The selected synset for the third comparison (synsets in English WN)

N	Arabic Word 1	Arabic Word 2	English Word 1	English Word 2	Synset Word 1	Synset Word 2	Human Rating
01	تَصْدِيق	سَاحِل	Coast	Endorsement	seashore.n.01	endorsement.n.05	0.01
02	خَيْط	ظُهْر	Noon	String	noon.n.01	string.n.01	0.01
03	مَشْي	موقد	Stove	Walk	stove.n.01	walk.n.01	0.01
04	خُضَار	عَبْد	Slave	Vegetable	slave.n.01	vegetable.n.01	0.04

05	قَرْيَة	بَسْمَة	Smile	Village	smile.n.01	village.n.02	0.05
06	مشفى	ساحر	Wizard	Infirmary	sorcerer.n.01	hospital.n.01	0.06
07	حَمَامَة	تَلّ	Hill	Pigeon	hill.n.01	pigeon.n.01	0.08
08	الْماس	كَأس	Glass	Diamond	glass.n.02	diamond.n.01	0.09
09	جَبَل	حَبْل	Cord	Mountain	cord.n.03	mountain.n.01	0.13
10	شَاطِئ	غابَة	Forest	Shore	forest.n.01	shore.n.01	0.21
11	شَيْخ	ضَرِيح	sepulcher	Sheikh	burial_chamber.n.01	sheik.n.01	0.22
12	مِخَدَّة	أَدَاة	Tool	Pillow	tool.n.01	pillow.n.01	0.25
13	جَبَل	سَاحِل	Coast	Mountain	seashore.n.01	mountain.n.01	0.27
14	قَدَح	أَدَاة	Tool	Tumbler	tool.n.01	tumbler.n.02	0.33
15	شَاطِئ	رِحْلَة	Journey	Shore	journey.n.01	shore.n.01	0.37
16	سَفَر	حَافِلَة	Coach	Travel	coach.n.01	travel.n.01	0.4
17	فرن	طعام	Food	Oven	food.n.02	oven.n.01	0.44
18	صِيَام	عِيد	Feast	Fasting	feast.n.02	fast.n.01	0.49
19	وَسِيلَة	حَافِلَة	Coach	Means	coach.n.01	means.n.02	0.52
20	أُخْت	فَتَاة	Girl	Sister	female_child.n.01	sister.n.01	0.6
21	جبل	تَلّ	Hill	Mountain	hill.n.01	mountain.n.01	0.65
22	شَيْخ	سَيِّد	Master	Sheikh	master.n.08	sheik.n.01	0.67
23	خُضَار	طَعام	Food	Vegetable	food.n.02	vegetable.n.01	0.69
24	جَارِيَة	عَبْد	Slave	Odalisque	slave.n.01	odalisque.n.01	0.71
25	مَشْي	جَرْي	Run	Walk	run.n.01	walk.n.01	0.75
26	خَيْط	حَبْل	Cord	String	cord.n.03	string.n.01	0.77
27	أَحْرَاش	غابَة	Forest	Woodland	forest.n.02	forest.n.02	0.79
28	مِخَدَّة	مِسْنَد	Cushion	Pillow	cushion.n.03	pillow.n.01	0.85
29	قَرْيَة	رِيف	Countryside	Village	countryside.n.01	village.n.02	0.85
30	شَاطِئ	سَاحِل	Coast	Shore	seashore.n.01	shore.n.01	0.89
31	وَسِيلَة	أَدَاة	Tool	Means	instrument.n.02	means.n.01	0.92
32	فَتَى	صَبِيّ	Boy	Lad	male_child.n.01	cub.n.02	0.93
33	قَبْر	ضَرِيح	Sepulcher	Grave	burial_chamber.n.01	grave.n.02	0.94
34	مشعوذ	ساحر	Wizard	Magician	sorcerer.n.01	sorcerer.n.01	0.94
35	قَدَح	كَأس	Glass	Tumbler	glass.n.02	tumbler.n.02	0.95

Table 11. Result of the third comparison (applying similarity measures on English WN)

N	HR	W	P	LC	R	L
01	0.01	0.13	0.07	1.00	0.00	0.00
02	0.01	0.11	0.06	0.80	0.00	0.00
03	0.01	0.09	0.05	0.59	0.00	0.00
04	0.04	0.33	0.11	1.44	0.80	0.09
05	0.05	0.13	0.07	1.00	0.00	0.00
06	0.06	0.44	0.09	1.24	1.53	0.15
07	0.08	0.32	0.07	1.00	1.29	0.13
08	0.09	0.56	0.11	1.44	2.31	0.23
09	0.13	0.40	0.10	1.34	1.29	0.12
10	0.21	0.18	0.10	1.34	0.00	0.00
11	0.22	0.42	0.09	1.24	1.53	0.13
12	0.25	0.63	0.14	1.69	2.31	0.25
13	0.27	0.67	0.20	2.03	5.88	0.60
14	0.33	0.71	0.17	1.85	3.26	0.32
15	0.37	0.13	0.07	1.00	0.00	0.00
16	0.4	0.13	0.07	0.93	0.00	0.00
17	0.44	0.24	0.07	1.00	0.80	0.09
18	0.49	0.47	0.10	1.34	2.04	0.18
19	0.52	0.47	0.10	1.34	1.53	0.17
20	0.6	0.60	0.14	1.69	2.33	0.25
21	0.65	0.83	0.33	2.54	6.95	0.73
22	0.67	0.63	0.16	1.84	2.33	0.20
23	0.69	0.83	0.33	2.54	6.11	0.84
24	0.71	0.60	0.14	1.69	2.33	0.00
25	0.75	0.57	0.10	1.34	3.73	0.42
26	0.77	0.59	0.13	1.56	2.31	0.20
27	0.79	1.00	1.00	3.64	9.61	1.00
28	0.85	0.93	0.50	2.94	11.29	0.98
29	0.85	0.75	0.20	2.03	4.76	0.42
30	0.89	0.91	0.50	2.94	9.42	0.96
31	0.92	0.93	0.50	2.94	6.79	0.83
32	0.93	0.95	0.50	2.94	8.40	0.83
33	0.94	0.93	0.50	2.94	9.85	0.96
34	0.94	1.00	1.00	3.64	11.98	1.00
35	0.95	0.94	0.50	2.94	9.44	0.81

The first row is represented by the following abbreviations: N: the numbering of the word, HN: Human rating, W: WuP measure, Li: P: Path measure, LC: Leakcock & Chodorow, R: Resnik measure and L: Lin.

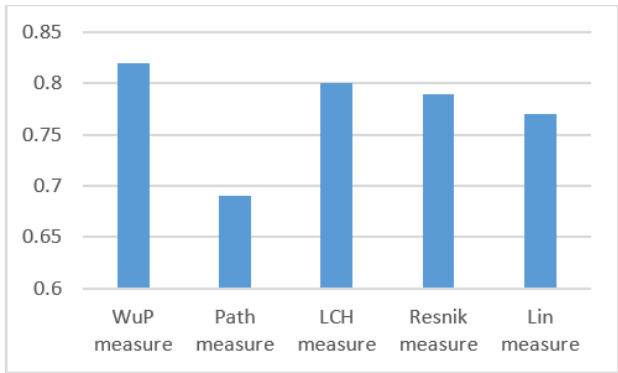

Figure 6. The correlation between human ratings and similarity measures scores in in AWN experiment using the correct synsets

According to Table 12, the WuP measure in English WN yielded the best similarity scores among all measures in experiments 2 and 3, as indicated by its highest correlation value (Figure 6) and lowest mean squared error.

Table 12. The correlation and MSE of the measures in the third comparison

Measures	Correlation	MSE
WuP measure	0.82	0.042
Path measure	0.69	0.117
LCH measure	0.80	0.041
Resnik measure	0.79	0.070
Lin measure	0.77	0.069

3.5 Experimental observations and the introduction of a novel hybrid Similarity measure combining WuP and Resnik measures

3.5.1 A comparison of AWN and WN

This section explores the contrasts between AWN and WN identified in previous experiments, along with the benefits and drawbacks of utilizing each of them.

Advantages of leveraging AWN in the domain of natural language processing for Arabic include:

- Every word in AWN is marked with Arabic vowels, making it better equipped for avoiding homonyms when seeking a synset for an Arabic word.

(1) It performs exceptionally well when processing Arabic books that contain Arabic vowels (diacritics).

(2) Outperforms other ontologies in tasks related to natural language processing, such as determining word roots.

Regarding the constraints associated with the utilization of AWN in the context of natural language processing for Arabic, these encompass:

- AWN's limited synset coverage requires WN to locate missing synsets, which are not available in AWN.
- Decreased efficiency in tasks involving words without Arabic vowels.
- AWN has not received an update since 2010.
- When used with modern books and online publications lacking Arabic vowels, it has poor efficiency.
- The lack of resources for the Arabic language limits the use of many similarity measures.

While the merits of employing WN in the realm of natural language processing for Arabic encompass:

- WN contains a larger number of synsets than other language specific WNs.
- WN has been updated more frequently than AWN.
- Provides better accuracy in calculating the degree of similarity between concepts.
- Allows for the use of all similarity measures.
- Performs better with modern documents because they lack Arabic vowels, only requiring word translation.
- English WN has a wealth of resources in terms of corpus and tools, making it easier to implement than other WNs.

Drawbacks associated with the utilization of WN in the context of natural language processing for Arabic comprise:

- Requires correct translation of words to obtain the desired result.
- Cannot be used in many Arabic language processing tasks.

3.5.2 Comparing results of similarity measurements and introduction of a hybrid measure

We analyzed the results of similarity calculation using various measures and compared the measures that yielded the best outcomes in experiments where we used the correct synset.

As previously mentioned, the WuP measure exhibited a high correlation coefficient (0.82) with human ratings, indicating a good linear relationship with human ratings. Figure 7 illustrates the squared error between human ratings and scores obtained from three different measures.

Through this experiment, we note that measure WuP has a good ability to measure the similarity between words, where we note that the results of measure WuP are excellent for words that have a large percentage of similarity, while we note relatively irregular values for words that have little similarity, unlike measure Resnik , which was able to get very close to the exact result of these words from the conclusion of the observation that we made in many experiments, it can be concluded that measure WuP can be used as a good separator between words that are close in meaning and words that are far apart in meaning.

From the Figure 7, it is evident that WuP outperformed the other measures on pairs with similarity scores above 0.55. On the other hand, for pairs with similarity scores below 0.55, the squared error was high, whereas the Resnik measure exhibited better performance on pairs with similarity scores below 0.45. Finally, the LCH measure performed well on pairs with moderate similarity scores (between 0.45 and 0.55).

According to the results of the similarity measures, it was found that the WuP measure is more effective in determining the similarity of word pairs with a similarity score greater than a certain threshold, while providing a slightly higher score for pairs with a similarity value between them below this threshold. Conversely, the Resnik measure is more suitable for calculating the similarity score between words with similarity values below this threshold (1).

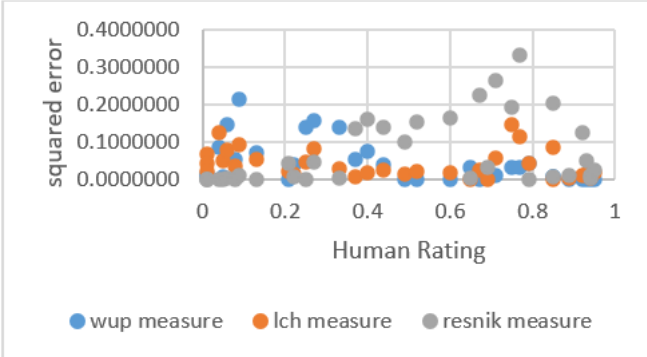

Figure 7. Comparison of similarity measures squared error values in experiment 3 (the result of applying measure on WN)

Our earlier experiments indicate that employing a sole measure might not adequately capture the accurate degree of similarity between two words. We advocate for a more intricate method in gauging the similarity between word pairs. Rather than depending on a single metric, we propose categorizing word pairs into 'semantically close' and 'not semantically close'. Each category requires a distinct measure, as our results demonstrate that a universal measure lacks efficacy. This approach is expected to produce.

Based on these observations, a new similarity measure is proposed that incorporates the WuP and Resnik measures. The proposed measure takes into account several factors, including the depth of concept, the length of the shortest path between the synsets of two words, and the information content of their Least Common Subsumer (LCS).

$$sim_{Resnik\ \&\ WuP}(c_1, c_2) = \begin{cases} sim_{resnik}(c_1, c_2), sim_{wup}(c_1, c_2) < threshold \\ sim_{resnik}(c_1, c_2), sim_{wup}(c_1, c_2) \geq threshold \end{cases} sim_{Resnik\ \&\ WuP}(c_1, c_2)$$

$$= \begin{cases} IC\big(lso(c_1, c_2)\big), \dfrac{2 \times depth\big(lso(c_1, c_2)\big)}{len(c_1, c_2) + 2 \times depth\big(lso(c_1, c_2)\big)} < threshold \\ \dfrac{2 \times depth\big(lso(c_1, c_2)\big)}{len(c_1, c_2) + 2 \times depth\big(lso(c_1, c_2)\big)}, \dfrac{2 \times depth\big(lso(c_1, c_2)\big)}{len(c_1, c_2) + 2 \times depth\big(lso(c_1, c_2)\big)} \geq threshold \end{cases} \quad \text{and threshold} = 0.45 \tag{1}$$

To compute the similarity value between two concepts, we first apply the WuP measure. If the resulting value exceeds a specific threshold, we consider it as the final result. Otherwise, we use the Resnik measure to obtain the final result. It is important to mention that the threshold value has been determined using the provided dataset, and we suggest it to be within the range of 0.4 to 0.46, with WuP producing excellent similarity values above this threshold. Nonetheless, it would be preferable to identify the optimal value based on a larger dataset.

3.5.3 Result of the hybrid measure

The results of applying the hybrid measure that combines WuP and Resnik measures are presented in Table 13.

Table 13. The results of the hybrid measure combining WuP and Resnik measures

N	Human Rating	Hybrid Measure	Error	Squared Error
01	0.01	0.00	0.01	0.0001
02	0.01	0.00	0.01	0.0001
03	0.01	0.00	0.01	0.0001
04	0.04	0.07	0.03	0.0007
05	0.05	0.00	0.05	0.0025
06	0.06	0.13	0.07	0.0046
07	0.08	0.11	0.03	0.0008
08	0.09	0.56	0.47	0.2167
09	0.13	0.11	0.02	0.0005
10	0.21	0.00	0.21	0.0441
11	0.22	0.13	0.09	0.0085
12	0.25	0.63	0.38	0.1406
13	0.27	0.67	0.40	0.1573
14	0.33	0.71	0.38	0.1413
15	0.37	0.00	0.37	0.1369
16	0.4	0.00	0.40	0.1600
17	0.44	0.07	0.37	0.1392
18	0.49	0.47	0.02	0.0004
19	0.52	0.47	0.05	0.0024
20	0.6	0.60	0.00	0.0000
21	0.65	0.83	0.18	0.0336
22	0.67	0.63	0.04	0.0016
23	0.69	0.83	0.14	0.0205
24	0.71	0.60	0.11	0.0121
25	0.75	0.57	0.18	0.0319
26	0.77	0.59	0.18	0.0330
27	0.79	1.00	0.21	0.0441
28	0.85	0.93	0.08	0.0069
29	0.85	0.75	0.10	0.0100
30	0.89	0.91	0.02	0.0004
31	0.92	0.93	0.01	0.0002
32	0.93	0.95	0.02	0.0003
33	0.94	0.93	0.01	0.0000
34	0.94	1.00	0.06	0.0036
35	0.95	0.94	0.01	0.0001
Correlation =		**0.85**	**MSE =**	**0.038**

The effectiveness of the Hybrid measure can be observed in Figure 8, particularly for word pairs with a high level of similarity. However, as with many other similarity measures, the efficiency of the measure decreases for word pairs with a moderate degree of similarity. Despite this, there is an improvement in the correlation between human ratings and the measure's output, with a correlation coefficient of 0.85 and a decrease in the Mean Squared Error to 0.038.

Figure 8. Squared error results of the hybrid similarity measure

4. CONCLUSIONS AND FUTURE WORKS

In conclusion, this research paper comprehensively explored various facets of semantic similarity measures in the context of Arabic natural language processing (NLP).

Our study conducted an in-depth comparison between Arabic WordNet (AWN) and WordNet (WN) in the domain of Arabic NLP. Our investigation revealed distinct advantages and limitations associated with both resources. AWN's notable strengths lie in its Arabic vowel markings, exceptional performance with diacritics-laden Arabic books, and proficiency in tasks involving word root determination. However, AWN's drawbacks include limited synset coverage, reduced efficiency with non-voweled words, lack of updates, and inefficacy with modern publications lacking Arabic vowels. The scarcity of resources further constrains the application of various similarity measures.

Our research delved into the comparison of similarity measurement results using different measures. Notably, the WuP measure exhibited a strong correlation with human ratings (0.82), particularly for word pairs with high similarity. Conversely, the Resnik measure showed better performance for word pairs with low similarity. These findings were instrumental in guiding the development of a novel hybrid similarity measure.

The significance of our study lies in its contribution to advancing the understanding of semantic similarity measures. We introduce a novel hybrid measure, combining the strengths of WuP and Resnik measures, and optimize its performance using empirical thresholding. This hybrid measure exhibits

substantial improvements in the correlation between human ratings and model output, boasting a remarkable correlation coefficient of 0.85 and a significantly reduced Mean Squared Error of 0.038.

In summary, our research not only comprehensively compared AWN and WN for Arabic NLP but also proposed an innovative hybrid similarity measure that significantly enhances semantic similarity calculations. This study thus makes a notable contribution to the field of NLP, underscoring the importance of resource selection and hybrid measures in achieving more accurate and context-aware semantic similarity assessments in the Arabic language. Nevertheless, it is essential to recognize that word similarity is not static, constrained solely by dictionary definitions. It mutates with evolving cultures and the emergence of new terminologies. This underscores the variances in similarity perceptions across distinct human groups, owing to prevailing cultural nuances. Consequently, we advocate the exploration of dynamic ontologies, subject to automatic evolution through artificial intelligence techniques. Such an approach holds the promise of enhancing the precision of word similarity assessments and, consequently, improving the efficacy of NLP applications.

REFERENCES

[1] Muzakir, A., Adi, K., Kusumaningrum, R. (2023). Advancements in semantic expansion techniques for short text classification and hate speech detection. Ingénierie des Systèmes d'Information, 28(3): 545-556. https://doi.org/10.18280/isi.280302

[2] Bhyrapuneni, S., Rajendran, A. (2022). Word recognition method using convolution deep learning approach used in smart cities for vehicle identification. Revue d'Intelligence Artificielle, 36(3): 489-495. https://doi.org/10.18280/ria.360318

[3] Lazarre, W., Guidedi, K., Amaria, S., Kolyang. (2022). Modular ontology design: A state-of-art of diseases ontology modeling and possible issue. Revue d'Intelligence Artificielle, 36(3): 497-501. https://doi.org/10.18280/ria.360319

[4] Gruber, T.R. (1995). Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies, 43(5-6): 907-928. https://doi.org/10.1006/IJHC.1995.1081.

[5] Kulmanov, M., Smaili, F.Z., Gao, X., Hoehndorf, R. (2021). Semantic similarity and machine learning with ontologies. Briefings in Bioinformatics, 1-18. https://doi.org/10.1093/BIB/BBAA199

[6] Sánchez, D., Batet, M., Isern, D., Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications, 39(9): 7718-7728. https://doi.org/10.1016/J.ESWA.2012.01.082

[7] Pawar, A., Mago, V. (2019). Challenging the boundaries of unsupervised learning for semantic similarity. IEEE Access, 7: 16291-16308. https://doi.org/10.1109/ACCESS.2019.2891692

[8] Zhu, G., Iglesias, C.A. (2017). Computing semantic similarity of concepts in knowledge graphs. IEEE Transactions on Knowledge and Data Engineering, 29(1): 72-85. https://doi.org/10.1109/TKDE.2016.2610428

[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] Elkateb, S., Black, W. J., Vossen, P., Farwell, D., Rodríguez, H., Pease, H., Alkhalifa, M., Fellbaum, C. (2006). Arabic WordNet and the challenges of Arabic. In Proceedings of the International Conference on the Challenge of Arabic for NLP/MT, 15-24. https://aclanthology.org/2006.bcs-1.2.

[11] Suggested Upper Merged Ontology (SUMO) - GM-RKB. (n.d.). http://www.gabormelli.com/RKB/Suggested_Upper_Merged_Ontology_(SUMO).

[12] Rada, R., Mili, H., Bicknell, E., Blettner, M. (1989). Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1): 17-30. https://doi.org/10.1109/21.24528

[13] Bulskov, H., Knappe, R., Andreasen, T. (2002). On measuring similarity for conceptual querying. In Flexible Query Answering Systems: 5th International Conference, Copenhagen, Denmark, pp. 100-111. https://doi.org/10.1007/3-540-36109-X_8

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

[15] Wu, Z., Palmer, M. (1994). Verb semantics and lexical selection. arXiv preprint cmp-lg/9406033, 133-138. https://doi.org/10.48550/arxiv.cmp-lg/9406033

[16] Almarsoomi, F.A., OShea, J.D., Bandar, Z., Crockett, K. (2013). AWSS: An algorithm for measuring Arabic word semantic similarity. In 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, pp. 504-509. https://doi.org/10.1109/SMC.2013.92

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

[18] Li, Y., Bandar, Z.A., Mclean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, 15(4): 871-882. https://doi.org/10.1109/TKDE.2003.1209005.

[19] Aldiery, M.G. (2017). The semantic similarity measures using Arabic ontology هباشتلا يللادلا سيياقم. Doctoral dissertation, Middle East University.

[20] Banu, A., Fatima, S.S., Khan, K.U.R. (2015). Information content based semantic similarity measure for concepts subsumed by multiple concepts. International Journal of Web Applications, 7(3): 85-94.

[21] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. arXiv Preprint CMP-lg/9511007.

[22] Saruladha, K., Aghila, G., Raj, S. (2010). A new semantic similarity metric for solving sparse data problem in ontology based information retrieval system. International Journal of Computer Science Issues, 7(3): 40-48.

[23] Sánchez, D., Batet, M., Isern, D. (2011). Ontology-based information content computation. Knowledge-Based Systems, 24(2): 297-303. https://doi.org/10.1016/j.knosys.2010.10.001

[24] Seddiqui, M.H., Aono, M. (2010). Metric of intrinsic information content for measuring semantic similarity in an ontology. In Proceedings of the Seventh Asia-Pacific

Conference on Conceptual Modelling, 110: 89-96.
[25] 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.
[26] Jiang, J.J., Conrath, D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv Preprint CMP-lg/9709008.
[27] Lin, D. (1998). An information-theoretic definition of similarity. In ICML, 98: 296-304.
[28] Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A. (2003). Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. Bioinformatics, 19(10): 1275-1283. https://doi.org/10.1093/bioinformatics/btg153
[29] Seco, N., Veale, T., Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In ECAI, 16: 1089-1090.
[30] Meng, L., Gu, J. (2012). A new method for calculating word sense similarity in WordNet1. International Journal of Signal Processing, Image Processing and Pattern Recognition, 5(3): 197-206. https://www.earticle.net/Article/A208828
[31] Chandrasekaran, D., Mago, V. (2021). Evolution of semantic similarity—A survey. ACM Computing Surveys (CSUR), 54(2): 1-37. https://doi.org/10.1145/3440755
[32] Tversky, A. (1977). Features of similarity. Psychological Review, 84(4): 327-352. https://doi.org/10.1037/0033-295X.84.4.327
[33] Ezzikouri, H., Madani, Y., Erritali, M., Oukessou, M. (2019). A new approach for calculating semantic similarity between words using WordNet and set theory. Procedia Computer Science, 151: 1261-1265. https://doi.org/10.1016/j.procs.2019.04.182
[34] Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, pp. 24-26. https://dl.acm.org/doi/pdf/10.1145/318723.318728
[35] Patwardhan, S., Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together.
[36] Patwardhan, S. (2003). Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Doctoral dissertation, University of Minnesota, Duluth.
[37] Merhbene, L., Zouaghi, A., Zrigui, M. (2013). A semi-supervised method for Arabic word sense disambiguation using a weighted directed graph. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, pp. 1027-1031.
[38] Jiang, Y., Zhang, X., Tang, Y., Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. Information Processing & Management, 51(3): 215-234. https://doi.org/10.1016/j.ipm.2015.01.001
[39] Zhou, Z., Wang, Y., Gu, J. (2008). New model of semantic similarity measuring in wordnet. In 2008 3rd International Conference on Intelligent System and Knowledge Engineering, Xiamen, China, 1: 256-261. https://doi.org/10.1109/ISKE.2008.4730937
[40] Faaza, A., James, D., Zuhair, A., Keeley, A. (2012). Arabic word semantic similarity. International Journal of Cognitive and Language Sciences, 6(10): 2497-2505. https://doi.org/10.5281/zenodo.1080052