Evaluating the Impact of Semantic Gaps on Estimating the Similarity Using Arabic Wordnet

Mamoun Abu Helou*

Management Information System Department, Al-Istiqlal University, Jericho, 11590, Palestine

ARTICLE INFO

Article history:
Received: 30 August, 2020
Accepted: 21 October, 2020
Online: 26 October, 2020

Keywords:
Lexical ontologies
Arabic wordnet
Semantic gaps
Semantic structure
Semantic similarity measure

ABSTRACT

Knowledge-based approach is widely used in various NLP applications. For example, to evaluate the semantic similarity between words, the semantic evidence in lexical ontologies (wordnets) is commonly used. The success of the English WordNet (EnWN) in this domain has inspired the creation of several wordnets in different languages, including the Arabic WordNet (ArWN). The English synsets have been extended to Arabic synsets through translation, which have introduced semantic gaps in ArWN structure. Therefore, compared to EnWN, ArWN has limited coverage in terms of lexical and semantic knowledge. This paper explores to what degree the richness of the wordnets’ semantic structure influences the semantic evidence that can be used in wordnet-based applications, in particular the effect of filling the semantic gaps in ArWN. The paper studies the performance of applying English-based and Arabic-based similarity measures over ArWN. A set of experiments was performed by applying six path-based semantic similarity measures over Arabic benchmark dataset to investigate the usability and efficacy of the enriched structure of ArWN. The Performance measures, Person Correlation and Mean Square Error, are computed against and compared to human judgment benchmark. The obtained results demonstrate that the semantic similarity between words can be significantly improved when filling the semantic gaps. In addition, the experiment findings show that Arabic-based measures competitively perform well compared to the English-based measures. Further, ArWN enhanced structure is also available for public.

1 Introduction

In Natural Language Processing applications, a common task is to estimate the semantic similarity among words [1]. Lexical resources, such as, bilingual and multilingual dictionaries, thesauruses, lexical ontologies (wordnets), machine translation services among others, are widely used to estimate the similarity [2]. For instance, various tasks of natural language processing, knowledge engineering, and computational linguists have exploited the lexical and semantic knowledge encoded in the English WordNet (EnWN) [3,4]; including sense disambiguation, information retrieval, text summarization, and question answering [5–6].

EnWN has been expanded to provide multilingual knowledge in many wordnet projects [7–8]. The Arabic WordNet (ArWN) [9] has extended EnWN by translating English synsets. However, English synsets that do not have translation in Arabic introduce semantic gaps in ArWN’s semantic structure. For instance, synsets containing a single and polysemous word are difficult to determine their meaning by means of direct translation; in fact, more evidence is required to disambiguate their meaning [10–11]. Thus, similarity measures designed for English (i.e., English-based similarity measures) may not be effective in the same way when applied over resources in other languages; in this work we consider Arabic language.

Experiment findings in [12] showed that ArWN has limited coverage of lexical and semantic knowledge compared to EnWN. Further attempts have been made to improve the content of ArWN [9, 13–14]. However, resolving the semantic gaps was not considered. In [15,16], they studied the performance of different similarity measures over ArWN. However, no explicit configuration was stated when calculating the similarity scores. Further, no explanation was given on how some semantic similarity scores were reported.

In [17], a preliminary study was conducted to examine the impact of the semantic gaps on estimating the semantic similarity scores using ArWN. They examined the impact of improving the semantic structure of ArWN on estimating the similarity between Ar-
bic synsets. The semantic gaps were analyzed and identified. Then new synsets in Arabic were added to ArWN and mapped to their corresponding synsets in English, using interactive cross-lingual mapping approach \cite{13}. The impact of the enriched ArWN was studied in semantic similarity experiment using only one English-based semantic similarity measure.

In this paper we extend previous work presented in \cite{17}; a large scale experiment is conducted to further examine the degree to which wordnet-based applications can be influenced by improving their semantic structure, mainly considering ArWN. In particular, the main contributions of this work can be summarized as follow:

(i) Four settings are defined and applied over two variants of ArWN structure. In the experiment six path-based similarity measures are applied over ArWN and EnWN; including, four English-based similarity measures (Path \cite{19}, Li \cite{2}, WuP \cite{20}, and Lch \cite{21}), and two Arabic-based similarity measures (AWSS \cite{22}, and Aldirey \cite{16}).

(ii) Study to which extent the semantic similarity measures that are developed for Arabic-based applications can perform efficiently well compared to English-based similarity measures. A comprehensive comparison between the similarity measures over the different configurations is provided, for both EnWN and ArWN.

The similarity scores obtained from the different measures, in the different settings, are compared to a standard benchmark for Arabic word pairs obtained from the AWSS dataset \cite{23}. Two measures, the Person Correlation and the Mean Square Error measures, are used to quantify the performance of the similarity measures. Reported values indicate the importance of the semantic evidence obtained from the enrichment process, and its significant effect on estimating the semantic similarity between words. In addition, the results show that Arabic-based measures performs competitively good compared to English-based measures.

The rest of this paper is organized as follows. Section 2 overviews related works on building wordnets, and the development of wordnet-based semantic similarity measures. Section 3 and describes the approach used to evaluate the impact of Semantic Gaps on estimating the Similarity over ArWN. Section 4 discusses experiments conducted: the benchmark dataset, the performance measures, and the obtained results. Finally Section 5 draws some conclusions and outlines future work.

2 Related works

This section provides an overview of the construction of wordnets and the ArWN contents; presents wordnet-based semantic similarity measures, which will be used in the experiment.

2.1 Wordnets overview

Wordnets, also known as lexical ontologies \cite{24}, are considered to be a resource of lexical and semantic knowledge, which organize natural language words (lexicons) into synsets. A synset is a collection of synonym words that express one meaning in a specific context (i.e., concept) \cite{5} \cite{25}.

In wordnets, words are arranged in a lexical database. Words can have several senses, such that each sense of a given word is identified by a number and its part of speech type. For instance, the sense village#n#2 indicates the second (#2) nominal (#n) sense of the word “village”. Words are linked through lexical relations, for example, antonym and synonym relations. When a word can have more than one meaning, it is called polysemous word, which can be member of several synsets. Otherwise, it is called monosemous word, which is a member of a single synset. For example, the word “village” has three noun senses as defined in EnWN; which are indicated in the following set of synsets:\{village#n#1, small-town#n#1, settlement#n#2\}, \{village#n#2, hamle#n#3\}, \{Greenwich-village#n#1, village#n#3\}.

Synsets are related by semantic relations. The Hypernymy and Hyponymy relations are considered to be the key semantic relations that form the semantic structure in wordnets. Hypernymy is described as the inverse of Hyponymy. For instance, in Figure 1 the synset \{village#n#2, hamle#n#3\} is hyponymy of the synset \{settlement#n#6\}, while the synset \{settlement#n#6\} is hyponymy of the synset \{village#n#2, hamle#n#3\}. Further, definitions (glosses) are also attached to synsets to convey their meaning. For example, the word sense village#n#2 defined as “a settlement smaller than a town” \cite{1}.

The HyperTree of a given synset (i.e, word sense) is defined as the sequence of synsets that are linked with hypernymy relations, which connect a synset with its ancestor synsets up to the root node. The function HyperTrees(word) produces the set of HyperTrees which a given word belongs. Figure 7 shows an excerpt of a HyperTrees in English and their correspondence in Arabic \cite{1}.

EnWN has been manually produced at Princeton University over the past three decades \cite{3} \cite{4}. EnWN’s success in many computational language domains has inspired the development of similarly structured lexicons, for both individual and multiple languages \cite{26}, such as EuroWordNet \cite{7}, BalkaNet \cite{27}, Polylingual WordNet \cite{8}, universal wordnet \cite{28}, MultiWordNet \cite{29}, WikiNet \cite{30}, and Arabic WordNet \cite{9}.

Computational linguistics has defined the Inter-Lingual Index \cite{7}, to establish links between different wordnets which is considered to be independent of language. For instance, near-equivalence and equivalence semantic relations are used to link synsets from the individual wordnets to the Inter-Lingual Index. Wordnets for several languages have been developed under the guidance of the Global WordNet Association \cite{1} which seeks to organize the creation and linking of wordnets. Further, the Open Multilingual WordNet project \cite{31} offers access to open wordnets in a number of languages, which are all connected to the latest version of EnWN (v3.0) \cite{1}.
2.2 Arabic wordnet contents

In the construction of ArWN [9], the extend method has been adopted. English Synsets have been translated into Arabic; and the structure of the EnWN (v2.0) has been inherited by ArWN. In the release of ArWN (v2.0) [1], 23,841 Arabic words, such as broken plurals, Named Entities, and roots have formed 11,296 synsets. Twenty-two types of semantic relationships have been used to connect synsets that formed 161,705 semantic links. Consequently, and in comparison with EnWN, which contains 147,306 words (117,659 synsets) [7], one can observe that ArWN has a limited coverage in terms of semantic relations and lexicons [12].

To this end, many attempts have been made to enhance the quality of ArWN by expanding its lexical coverage [13, 9] or semantic relationships [22] by different approaches. In [22], they released their work under the Lexical Markup Framework. However, the public release of ArWN ignores the synsets that are not linked to EnWN [21]. Nevertheless, synsets (semantic gaps) which are resolved in this work will be made for public [4]. In future work we plan to compile an xml format of ArWN enhanced structure, to enable researcher to utilize the ArWN in different applications.

2.3 Wordnet-based similarity measures

In linguistics, philosophy and information theory, estimating the semantic similarity between concepts is extensively studied [2, 15], which is a common and crucial task in many NLP applications, text summarization, word sense disambiguation, entailment, machine translation, among many others [33-36, 34, 35].

Estimating the semantic similarity between words is calculated by measuring the similarity between concepts (synsets) associated with the words [2]. Given two words, one can calculate the semantic similarity by exploiting wordnet (i.e., a lexical knowledge base). The lexical and semantic knowledge in wordnet have been used in many semantic similarity measures, which are originally designed and evaluated over EnWN (English-based measures) [36, 2].

In [15] they defined four broad categories of the similarity measures; Path-based similarity measures [21], Feature-based similarity measures [39], Information content similarity measures [37, 38], and Hybrid similarity measures [40, 41]. There have been few works concerned with the similarity of Arabic; AWSS measure [22] and Aldieri measure [16]. These have mainly adapted measures from those constructed for English. In particular, Li measure [2] was adapted, which is a path-based measure that consider the depth of concepts in the HyperTrees; the distance between two compared concepts; and the depth of the least common concept (lsc) that subsumed two compared concepts. Noting that, these measures needs to tune weighting parameters to find the optimal values [22, 16]. In this regards, several preliminary experiments are necessary to find the best weights that provide the optimal values.

An attempt to investigate the performance of the similarity measures over ArWN was conducted in [15]. They studied the performance of seven measures; including AWSS measure [22]. All measures were applied over 40 word pairs that are selected from AWSS dataset [23], which are also considered as the benchmark dataset in this work. The experiments findings [15] showed that

---

3ArWN obtained from the AWN browser available at: AWNBrowser, or as XML file available at: Open Multilingual Wordnet
4EnWN-3.0 available at: https://wordnet.princeton.edu/download
5Full list of the Semantic Gaps available at https://fas.alistiqlal.edu.ps/cv-2-ar.html

---

www.astesj.com
WuP measure [20] has the best performance in estimating the semantic similarity between Arabic word pairs. The experiments in [16] also introduced a competitive Arabic-based similarity measures (Aldiery measure) in comparison to WuP measure.

In [17] they further studied the impact of enhancing the HyperTree over the WuP measure. This work adopted and extend their experimental configurations and examine further the impact of the enhanced semantic structure of ArWN over Six measures including English and Arabic path-based measure, further details are provided in Section 4.

Recall that, for a given concepts \( c_i \) and \( c_j \), the function \( S_{\text{WuP}}(c_i, c_j) \) calculates the semantic similarity between \( c_i \) and \( c_j \), where \( m \) indicates the name of the measure. Next the description of the measures used in the experiment is given.

1. **Path measure** [19] finds the shortest path between the two concepts, by counting the number of edge (hypernymy relation) between the concepts, in order to compute the semantic similarity. Path measure which is considered as the pioneer similarity measure is defined in equation (1).

\[
sim_{\text{path}}(c_i, c_j) = \frac{1}{\text{len}(c_i, c_j)}
\]

Where the length function, \( \text{len}(c_i, c_j) \), returns the length of the shortest path between \( c_i \) and \( c_j \) in the wordnet semantic hierarchy. For example, in Figure 1, \( \text{len}(\text{hill#2, mountain#1}) = 3 \), and \( S_{\text{path}}(\text{hill#2, mountain#1}) = 0.333 \).

2. **Wup measure** [20] calculates the similarity by computing the distance between the two concepts and the maximum depth of the least common concept \( \text{lsc}(c_i, c_j) \) that subsumed the two concepts under evaluation. WuP measure is defined in equation (2).

\[
sim_{\text{WuP}}(c_i, c_j) = \frac{2 \times d(\text{lcs}(c_i, c_j))}{d(c_i) + d(c_j)}
\]

Where \( d(c_i) \) is the depth of the concept \( c_i \) using edge counting in the semantic hierarchy, \( \text{lcs}(c_i, c_j) \) is the least common subsumer of \( c_i \) and \( c_j \), \( d(\text{lcs}(c_i, c_j)) \) is the maximum length between \( \text{lcs} \) of \( c_i \) and \( c_j \) and the root of the hierarchy, where \( d(\text{entity}) = 1 \). For example in Figure 1, \( d(\text{hill#2}) = 7 \), \( d(\text{mountain#1}) = 7 \), \( d(\text{lcs(hill#2, mountain#1}) = 6 \), and \( S_{\text{WuP}}(\text{hill#2, mountain#1}) = 0.857 \).

3. **Lch measure** [21] uses the length of the shortest path between the two concepts, and also the maximum depth of the semantic hierarchy of a given part of speech type. Lch measure is defined in equation (3).

\[
sim_{\text{Lch}}(c_i, c_j) = -\log \frac{\text{len}(c_i, c_j)}{2 \times \text{maxDepth}_{\text{pos}}}
\]

Where, \( \text{maxDepth}_{\text{pos}} \) is the maximum depth of the hypernymy structure for a given part of speech. For instance, \( \text{maxDepth}_{\text{pos}} = 20 \) and 15 in EnWN and ArWN, respectively. For example in Figure 1, \( S_{\text{Lch}}(\text{hill#2, mountain#1}) = -\log(3/2 \times 20) = 2.590 \).

Noting that the Lch scores reported in Section 4 are normalized into the range 0 to 1 by dividing Lch scores over 3,688. Hence, \( S_{\text{Lch}}(\text{hill#2, mountain#1}) = 0.702 \).

4. **Li measure** [2] computes the similarity using non-linear function, which consumes the shortest length between concepts and the minimum depth of the concepts in the semantic hierarchy. Li measure is defined in equation (4).

\[
sim_{\text{Li}}(c_i, c_j) = e^{\alpha \text{len}(c_i, c_j)} \frac{e^{\beta d(\text{lcs}(c_i, c_j))} - e^{-\beta d(\text{lcs}(c_i, c_j))}}{e^{\beta d(\text{lcs}(c_i, c_j))} + e^{-\beta d(\text{lcs}(c_i, c_j))}}
\]

Noting that the parameters \( \alpha \) and \( \beta \) need to be calculated manually for good performance. The optimal parameters are \( \alpha = 0.2 \) and \( \beta = 0.6 \) as reported in [2]. For example, \( S_{\text{Li}}(\text{hill#2, mountain#1}) = 0.548 \).

5. **AWSS measure** [22] is an Arabic-based measure that adapted Li measure to compute semantic similarity with modification on the depth and length computation to be proper for ArWN [23]. AWSS measure is defined in equation (5).

\[
sim_{\text{AWSS}}(c_i, c_j) = e^{-\alpha d(\text{lcs}(c_i, c_j))} \times \text{tanh}(\beta \times \text{len}(c_i, c_j))
\]

Where the parameters \( \alpha \) and \( \beta \) are the length and depth factors respectively. The optimal performance was obtained at \( \alpha = 0.162 \) and \( \beta = 0.234 \) as reported in [22]. For example in Figure 1, \( \text{len}(\text{rukaAm}_1, \text{Jabal}_1) = 4 \) and \( d(\text{lcs}(\text{rukaAm}_1, \text{Jabal}_1)) = 8 \), then \( S_{\text{AWSS}}(\text{rukaAm}_1, \text{Jabal}_1) = 0.201 \).
6. Aldiery measure [16] is an Arabic-based measure also adapted Li measure to compute semantic similarity with modification on the depth and length computation to be proper for ArWN. Aldiery measure is defined in equation (6).

\[
\text{sim}_{\text{Aldiery}}(c_i, c_j) = \tanh\left(\frac{2 \cdot d(lsc(c_i, c_j))}{d(c_i) + d(c_j)}\right) + W\left(\frac{1}{\log(len(c_i, c_j))} + \frac{\log(len(c_i, c_j))}{\log(maxDepth_{pos})}\right)
\]

(6)

Where [16] defines \( W = 0.5 \). For example in Figure 1 \( d(\text{rukaAm}, \text{Jabal}) = 8 \), \( d(\text{Jabal}, \text{Jabal}) = 7 \), \( len(\text{rukaAm}, \text{Jabal}) = 4 \), and \( d(lsc(\text{rukaAm}, \text{Jabal}), \text{Jabal}) = 8 \), and \( maxDepth_{rukaAm} = 15 \), then \( \text{sim}_{\text{Aldiery}}(\text{rukaAm}, \text{Jabal}) = 0.692 \).

Noting that, the similarity functions defined above consume either words, or word senses as parameters. In the first case, the similarity function returns the highest similarity score for all the possible combination of word senses for the two given words. In the second case, it returns the similarity score between the two defined senses.

In addition, the six measures defined in the equations [12][14][15][16] and [6] are path-based measures, this study focus on the impact of the structure without interference of other semantic evidence such as features extracted from corporuses, which depend on the quality of the used cuprous, as well as the availability of resources in Arabic.

On the other hand, Path, Wup, and Lch measures are considered as linear path-based measures, while Li measure is a non-linear path based measure. AWSS and Aldiery are also non-linear path-based measures, which are derived from Li and purposely developed for Arabic.

Observe that, for the Path, Wup, and Lch measures no weights are required to be tuned. While the other measures need to find optimal value of the defined weights. The four English-based measures, as well as the two Arabic-based measures are selected because they achieved good performance against other measures [22][15], and to compare the performance between the measures using Arabic benchmark dataset.

3 Evaluating the impact of semantic gaps on estimating the similarity

This section presents the approach that is used to evaluate the impact of enhancing the structure of ArWN on estimating the semantic similarity. Figure 2 illustrates the main phases of the approach, which are explained as follow.

1. Synset Analysis. In this phase the semantic gaps are identified through a comparison between the structures of ArWN (v2.0) and EnWN (v3.0). In particular, for each nominal synset in ArWN, Hypernymy relations are compared with their EnWN correspondences. The HyperTrees for each synset in ArWN is compared with its correspondence HyperTrees in EnWN. For example, Figure 1 indicates two semantic gaps in the ArWN HyperTree(ŠaATi), HyperTree(AlbaHor).

2. HyperTrees Improvement. In this phase ICLM Web application [13] is used to fill the identified semantic gaps. ICLM is a semi-automatic matching approach that supports feedback provided by multiple users. In ICLM the number of users that are asked to perform each mapping task is estimated based on the lexical characterization of concepts under evaluation, i.e., on the estimation of the ambiguity conveyed by the concepts involved in mappings [42], with the assumption that as the selection tasks difficulties increase, more users agreement is required.

The candidate matching of the source concepts in Arabic are automatically computed to the English target concepts using a lexical based disambiguation algorithm [43]. The study [42] recommended that combining lexical resources improves the quality of translations and provide a valuable support for candidate match retrieval in cross-lingual ontology matching problems. Accordingly, translations of the missing synsets are collected by combining lexical knowledge from different external resources. English synset translation was
collected from; Google Translate, BabelNet, and Almaany dictionary.

The difficulties of the mapping selection tasks, that is determining the number of user which are asked to perform the task, are estimated using lexical characteristics of concepts under evaluation: Ambiguity of lexicalization, Synonymy, richness, and Uncertainty in the selection Step. The mapping tasks are validated by some users based on a CAUTIOUS strategy. The task difficulty level is estimated as Low, Mid, and High level. One, three, or five users are asked to perform the Low, Mid, or High tasks, respectively.

In [17] ten users (bilingual speakers) are asked to validate the mapping tasks, that is, to fill a semantic gap in ArWN, and accordingly define new link with EnWN, hence, import the semantic relations among the concepts. The top ten frequent semantic gap are listed in Table 2. As a result 94% of the identified gaps are resolved, that is more than 98% of HyperTrees are filled in.

Observe that, some concepts are hard to resolve, and more evidences are needed. For Example, {mechanism#3}, {attache#1}, and {climber#1} synsets, which contain a single and polysemous word, are hard to determine their meaning with direct translation and no context [42], for this reason in the validation task users did not reach an agreement. Noting that, the semantic gaps for every word sense in the benchmark dataset used in the experiment are resolved.

3. Calculate Similarity. In this phase similarity measures defined in Section 2.3 are applied over the ArWN and EnWN using Arabic benchmark dataset (AWSS dataset [23]). Different configuration explained in Section 4.4 are applied to calculate the semantic similarity using the WS4J online application (see Section 4.1). Similarity scores are reported and passed to the next phase.

4. Performance Evaluation. In this phase the obtained similarity scores are compared with Human Rating benchmark [23] using two performance measures; The Person Correlation measure (r) and the Mean Squared Error (MS E). Further details are provided in the experiment Section 4.3.

Table 2: Top ten Frequent Semantic Gaps in ArWN with EnWN correspondence synsets

| EnWN synset | Freq | Semantic gaps in ArWN |
|-------------|------|-----------------------|
| 1           | 4,525| (أكاسيد مغني) |
| 2           | 100  | (جماع، جمع) |
| 3           | 50   | (객등 검찰관) |
| 4           | 48   | (أكاسيد مغني) |
| 5           | 30   | (أكاسيد مغني) |
| 6           | 36   | (أكاسيد مغني) |
| 7           | 30   | (أكاسيد مغني) |
| 8           | 24   | (أكاسيد مغني) |

4 Experiment

The conducted experiment aims at studying the efficacy of the semantic evidence in ArWN. In particular, the experiment focuses on the improvement of hypernymy relations in the semantic structure of ArWN. The experiment studies the extent to which the semantic structure of ArWN affects measuring the semantic similarity between concepts. This section reports and discusses the results obtained from running a set of configurations for measuring the semantic similarity scores over ArWN and EnWN.

Next sections present the tool which is used to calculate the semantic similarity scores, the benchmark dataset, the measures used to evaluate the performance of the structure improvement, and discuss obtained results.

4.1 Similarity Measure Tools

Significant efforts are being made in developing similarity measures to consume ArWN content. For example, the Java ArWN API [12] The application consumes Arabic words with diacritics (vocalized), whereas the benchmark dataset in this experiment contains unvocalized (without diacritics) word pairs. If Arabic words are vocalized, similar to the work done in [16] [15], then their senses will be defined in advance. The experiment’s configuration DS (see Section 4.4) studies the performance of determining the word senses on the similarity scores.

To avoid predefined senses, in this experiment the similarity scores are obtained using the WS4J online application [13]. In computing the scores, WS4J uses EnWN’s semantic structure (v3.0), which is used to measure the similarity scores between Arabic words. Noting that, in this experiment Arabic senses under evaluation have the same structure of their correspondence senses in English, as the semantic gaps in ArWN has been improved and linked to EnWN (v3.0). The similarity scores between the Arabic concepts are then measured using their correspondence concepts in EnWN. In addition, WS4J provides the description of all HyperTree of words under evaluation. The HyperTrees which returned for EnWN are validated to obtain Arabic words’ HyperTrees with semantic gaps as depicted in Figure 1. For instance, this information is necessary to measure the similarity scores in uHT configuration,
4.2 Benchmark dataset

Similar to the work performed in [15, 16], the AWSS benchmark [22] will be used in this experiment. The obtained similarity scores will be compared with Human Judgments obtained from the dataset of AWSS [23]. The AWSS dataset contains 70 nominal word pairs of Arabic, divided into three similarity levels, Low, Medium, and High; 40 word pairs are selected and used in this experiment, listed in Table 3 which are also used in [15, 16].

| NO. | Sim. level | En Word Pairs | Ar Word Pairs | HR |
|-----|------------|---------------|---------------|----|
| 1   | Low        | Coast         | Endorsement   | 0.92 |
| 2   | low        | Noon          | String        | 0.26 |
| 3   | low        | Stove         | Walk          | 0.01 |
| 4   | low        | Cord          | midaay        | 0.06 |
| 5   | low        | Signature     | String        | 0.02 |
| 6   | low        | Boy           | Endorsement   | 0.03 |
| 7   | low        | Boy           | Midday        | 0.04 |
| 8   | low        | Smile         | Village       | 0.05 |
| 9   | low        | Noon          | Fasting       | 0.07 |
| 10  | low        | Glass         | Diamond       | 0.09 |
| 11  | low        | Sepulcher     | Sheikh        | 0.22 |
| 12  | low        | Countryside   | Vegetable     | 0.31 |
| 13  | mid        | Tumbler       | Tool          | 0.33 |
| 14  | mid        | Laugh         | Feast         | 0.34 |
| 15  | mid        | Girl          | Odalesqie     | 0.49 |
| 16  | mid        | Feast         | Fasting       | 0.49 |
| 17  | mid        | Coach         | Means         | 0.52 |
| 18  | mid        | Sage          | Sheikh        | 0.56 |
| 19  | mid        | Girl          | Sister        | 0.60 |
| 20  | mid        | Hen           | Pigeon        | 0.65 |
| 21  | mid        | Hill           | Mountain      | 0.65 |
| 22  | mid        | Master        | Sheikh        | 0.67 |
| 23  | mid        | Food           | Vegetable     | 0.69 |
| 24  | mid        | Slave         | Odalesqie     | 0.71 |
| 25  | mid        | Run           | Walk          | 0.75 |
| 26  | high       | Coast         | String        | 0.77 |
| 27  | high       | Forest        | Woodland      | 0.79 |
| 28  | high       | Sage          | Thinker       | 0.82 |
| 29  | high       | Journey       | Travel        | 0.84 |
| 30  | high       | Gem           | Diamond       | 0.84 |
| 31  | high       | Countryside   | Village       | 0.85 |
| 32  | high       | Cushion       | Pillow        | 0.85 |
| 33  | high       | Smile         | Laugh         | 0.87 |
| 34  | high       | Signature     | Endorsement   | 0.89 |
| 35  | high       | Tools         | Means         | 0.92 |
| 36  | high       | Sepulcher     | Grave         | 0.93 |
| 37  | high       | Boy           | Ladi          | 0.93 |
| 38  | high       | Wizard        | Magician      | 0.93 |
| 39  | high       | Coach         | Bus           | 0.95 |
| 40  | high       | Glass         | Tumbler       | 0.95 |

4.3 Performance Measures

The obtained similarity scores are evaluated against human ratings benchmark (HR), which is a human judgment similarity scores of Arabic nominal word pairs obtained from the dataset of AWSS.

Two measures are used to quantify the performance of the obtained similarity scores. The Person Correlation measure ($r$) defines the strength of the linear relationship between the obtained similarity scores and HR; the Mean Squared Error (MSE) calculates the average squared difference between the similarity scores and HR. The best performance is indicated by a similarity measure with the smallest MSE value and $r$ value is close to 1. While the negative $r$ value means that the obtained scores are increase as the HR ratings decrease. In addition, the similarity scores are compared to the performance results reported in [15, 16], which are listed in Table 4.

| Measure | $r$ | MSE |
|---------|-----|-----|
| 1       | 0.94 | 0.01648 |
| 2       | 0.89 | 0.0708 |
| 3       | 0.75 | 0.010038 |
| 4       | 0.85 | 0.02025 |
| 5       | 0.88 | 0.04424 |
| 6       | 0.96 | 0.01893 |

4.4 Experimental settings

Six path-based semantic similarity measures, which are defined in equations (1,2,3,4,5, and 6), will be applied over the Arabic word pairs benchmark dataset, which is described in Section 4.2. Using the following configurations, the similarity measures are applied over ArWN and EnWN to quantify the efficiency of ArWN structure enrichment:

1. **UnDefined Senses (uDS):** calculates the semantic similarity between given words without determining their senses. In this setting, which is considered as the default setting of the similarity measures, the similarity measure returns the maximum score obtained from the all possible combination of the senses of the given words.

2. **Defined Senses (DS):** calculates the semantic similarity between given words senses (i.e, sense are determined in advance). By extending the work in [17], the sense of each word pairs under evaluation is determined based on a majority vote (consensus) approach. Similar to the tasks of filling the semantic gaps [17, 18] (see Section 3), the CAUTIOUS strategy is adopted, where users are avoided to decide among word pairs that share the same words.

3. **wordnets Translation (wnTrans):** calculates the semantic similarity over ArWN by selecting the senses that match the
translations defined in the benchmark dataset. In *wnTrans* the maximum similarity score is selected, such that the ArWN and the EnWN cover the Arabic word and its translation in English, respectively. Otherwise, the default setting is *udS* is applied.

4. **Upper Bound (UB):** calculates the semantic similarity between given words senses, such that, *UB* selects the sense pair that maximize correlation values and minimize *MSE* values w.r.t the HR ratings (benchmark dataset). *UB* indicates the optimal scores for the considered experiment settings.

5. **Unimproved HyperTrees (uHT):** calculates the semantic similarity using ArWN while ignoring the structure enhancement. That is, the semantic gaps are considered in calculating the similarly scores.

6. **Improved HyperTrees (iHT):** calculates the semantic similarity using the enhanced structure of ArWN.

### 4.5 Results & Discussion

Tables 5, 6, 7, and 8 report the semantic similarity scores using six similarity measures, which resulted from applying *udS*, *DS*, *wnTrans* and *UB* configurations over ArWN; respectively. Such that two variants, *iHT* and *HT*, are considered. The tables also list the Arabic senses and their correspondences senses in English, which are used to provide the obtained similarity scores. Table 9 reports the semantic similarity scores that are obtained from applying *udS*, *DS*, and *UB* configurations over EnWN. English-based
between ArWn and EnWn; and the words their translation are
(signature: توقيع; sepulcher: جنازة; sage: حكيم; thinker: حكيم; pillow: مخدة; signature: توقيع; lad: لد). For example, the word “tumbler” has one sense in ArWn “tawoqiyE_l”, which is mapped into the “endorsement#5” in EnWn, while none of the five senses for the word “signature” in EnWn is mapped into ArWn. Noting that 28 word pairs out of the 40 word pairs has at least one missing correspondence sense in EnWn when considering uHT setting. For example: similarity scores of the 21st word pairs (hill “تله”; mountain “جبال”); which is also illustrated in
The performance measures $r$ and $MS$ are reported for every configuration in the bottom of Tables 6, 7, and 8; including the performance for each similarity level. Observe that, $r$ values show that $iHT$ achieves better performance compared to $uHT$. While, $MS$ values indicate that $uHT$ has less difference in similarity scores than $iHT$, compared to $HR$ rates. In fact, the values of $MS$ are strongly influenced by $uHT$, the semantic gaps. Noting that when HyperTrees of two senses have the same semantic gaps; the $lcs$ is reduced which decreases the similarity scores. This gives less difference in similarity scores compared to $HR$ rates. In particular, this happens for $MS$ values at mid similarity level. For example, in row 10, the HyperTrees of the word pairs (Glass; Diamond) has the (physical entity) as a semantic gap. That is, $d(glass#2) = 9$; $d(diamond#2) = 8$ and $d(lcs(glass#2, diamond#2)) = 3$; while $d(kwub#1) = 8$; $d(AlomAs#1) = 7$; $d(lcs(kwub#1, AlomAs#1)) = 2$.

Furthermore, wuTrans configuration scored the worst performance; this is due to the low Arabic word coverage. A significant finding is that, the richness of ArWN content has a high effect on the evaluation the semantic similarity between the concepts, in terms of the coverage of lexical and semantic relations.

Performance measures in Table 4 showed that WuP measure scored the best $MS$ value 0.0165 with 0.94 for $r$; and comparatively Alderiy measure has obtained the values 0.96
Overall, the reported performance values show that the enhancement of the semantic structure has a strong effect on estimating the semantic similarity between the concepts. Observe that, word pairs at low and mid similarity levels gives better r values than high similarity level. While words pairs in high similarity level gives better MSE values. In other words, similarity measures obtained best coloration values when the concepts are not similar. Both ArWN and EnWN, r and MSE measures indicate that best performance is achieved when word senses are determined in advance, i.e., DS configuration. However, it is important to distinguish the approach which is used to define the sense, in this work consensus based approach is used.

In other hand; the user feedback based approach, ICLM application that adopted to fill the semantic gaps, shows its effectiveness in selecting the senses, such that scores obtained in DS are close to optimal scores achieved with upper bound setting UB. Further, Arabic-based measure Aldiery performs better than AWSS, also

and 0.0189 for r and MSE, respectively. Nevertheless, [15] did not explicitly state which configuration was considered in calculating the similarity scores. For instance, in Table 8, the UB scores indicate the best value for r is 0.945 (with 0.542 for MSE) which is obtained by Aldiery measure; and the best MSE value is 0.0203 (with 0.935 for r) which is obtained by WuP measure. Further, in [16], semantic similarity scores were reported to be equal to zero for the word pairs in rows 1–9, which are at the low similarity level, and the word pair in row 21 was considered as not covered ArWN, hence, this increased the r values and reduced MSE values. However, no explanation is provided.
Table 9: uDS, DS, and UB configuration over EnWN

Aldiery measure provided a competitive performance in comparison to WuP measures.

5 Conclusion & Future Work

Six path-based similarity measures including English and Arabic based measures are applied over ArWN and EnWN to examine the effect of the improvement of the lexical and semantic coverage on wordnet-based similarity measures. wo variants uHT and iHT of ArWN structure are considered in the experiment to evaluate the impact of filling the semantic gaps on estimating the semantic similarity. The efficacy of the improved structure is examined by experiments in the context of semantic similarity. The semantic similarity scores for a benchmark dataset, human rating for 40 Arabic nominal word pairs, are calculated over ArWN and EnWN in different configurations (uDS, DS, wnTrans, and UB). The obtained performance values indicate the importance of the semantic evidence gained with the enrichment process; and its signification effect on estimating the semantic similarity between concepts. Moreover, when considering Arabic-based measures the experiment results showed that Aldiery measure performs better than AWS$^S$ measure. Beside that, Aldiery measure has provided a competitive performance in comparison to the English-based WuP measures. Finally, the resolved semantic gaps of the new structure are made for public.

As a future direction, we plan to compile xml format of the new structure, and to integrate it with available ArWN resources (i.e., ArWN release available at Open Multilingual WordNet [31]). It is also interesting to study the effect of the semantic gaps over NLP applications; for instances Question Answering similar to the work presented in [44], and word sense disambiguation [33, 35] in the context of Arabic.

References

[1] D. Jurafsky, J. H. Martin, Speech and Language Processing (2Nd Edition), Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.
[2] Y. Li, Z. A. Bandar, D. Mclean, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources,” IEEE Trans. on Knowl. and Data Eng., 15(4), 871–882, 2003, doi:10.1109/TKDE.2003.1209005.
[3] G. A. Miller, “WordNet: A Lexical Database for English,” Commun. ACM, 38(11), 39–41, 1995, doi:10.1145/219717.219748.
[4] C. Fellbaum, editor, WordNet An Electronic Lexical Database, The MIT Press, Cambridge, MA ; London, 1998.
[5] F. Christiane, H. Amanda, “When WordNet Met Ontology,” in Ontology Makes Sense - Essays in honor of Nicola Guarino, 136–151, 2019, doi: 10.3235/978-1-61499-955-3-136.

[6] N. Bouhriz, F. Benabbou, E. H. B. Lahmar, “Word Sense Disambiguation Approach for Arabic Text,” (IJACAS) International Journal of Advanced Computer Science and Applications, 7(4), 2016, doi:10.14569/IJACAS.2016.070451.

[7] P. Vossen, “Eurowordnet: A Multilingual Database Of Autonomous And Language-Specific Wordnets Connected Via An Inter-Linguinald,” International Journal of Lexicography, 17(2), 161–173, 2004, doi:10.1093/ijl/17.2.161.

[8] M. Arcan, J. P. McCrae, P. Buitelaar, “Polylingual Wordnet,” CoRR, abs/1903.01411, 2019.

[9] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. A. Martí, W. Black, S. Elkateb, J. Kirk, P. Vossen, C. Fellbaum, “Arabic WordNet: Current State and Future Extensions,” in Proceedings of the 4th International Conference on Global WordNet, 2008.

[10] G. A. Miller, W. G. Charles, “Contextual correlates of semantic similarity,” Language & Cognitive Processes, 6(1), 1–28, 1991.

[11] F. Bond, L. Morgado da Costa, M. W. Goodman, J. P. McCrae, A. Lohk, “Some Issues with Building a Multilingual Wordnet,” in Proceedings of the 12th Language Resources and Evaluation Conference, 3189–3197, European Language Resources Association, Marseille, France, 2020.

[12] L. Abouenour, K. Bouzoubaa, P. Rosso, “On the evaluation and improvement of Arabic Wordnet coverage and usability,” Language Resources and Evaluation, 47, 2013, doi:10.1007/s10579-013-9254-z.

[13] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. Alkalifa, A. Martí, “Arabic WordNet: Semi-automatic Extensions using Bayesian Inference,” 2008.

[14] M. A. Battita, M. Zrigo, “The Enrichment of Arabic WordNet Antonym Relations,” in A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, 342–353, Springer International Publishing, Cham, 2018, doi:10.1007/978-3-319-77113-7_27.

[15] N. Mohammed, D. Mohammed, “Experimental Study of Semantic Similarity Measures on Arabic WordNet,” International Journal of Computer Science and Network Security, 17(2), 2017.

[16] M. G. Aldayrly, The semantic similarity measures using Arabic ontology, (Master's thesis Theses and Dissertations Master). Middle East University, Jordan, 2017.

[17] M. A. Helou, “Effects of Semantic Gaps on Arabic WordNet-Based Similarity Measures,” in 2019 International Conference on Innovative Computing (ICIC), 1–10, 2019, doi:10.1109/ICIC4894.2019.8966672.

[18] M. A. Helou, M. Palmonari, “Multi-user Feedback for Large-scale Cross-lingual Ontology Matching,” in Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Funchal, Madeira, Portugal, November 1-3, 57–66, 2017, doi:10.5220/0006503200570066.

[19] R. Rada, H. Mili, E. Bicknell, M. Blettner, “Development and application of a metric on semantic nets,” in IEEE Transactions on Systems, Man and Cybernetics, 17–30, 1989, doi:10.1109/21.24528.

[20] Z. Wu, M. Palmer, “Verbs semantics and lexical selection,” in Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 133–138, Association for Computational Linguistics, Morristown, NJ, USA, 1994.

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

[22] F. A. Almarsoom, J. D. O'Shea, Z. Bandar, K. Crockett, “AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity,” in 2013 IEEE International Conference on Systems, Man, and Cybernetics, 504–509, 2013, doi:10.1109/SMC.2013.92.

[23] F. A. Almarsoom, J. D. O’Shea, Z. A. Bandar, K. A. Crockett, “Arabic Word Semantic Similarity,” International Journal of Cognitive and Language Sciences, 6(10), 2497 – 2505, 2012, doi:10.5281/zenodo.1080052.

[24] G. Hirst, “Ontology and the Lexicon,” in eds. S. Staab, R. Studer, editors, Handbook on Ontologies and Information Systems, Heidelberg, Springer, 2004, doi:https://doi.org/10.1007/978-3-540-24750-0_11.

[25] G. A. Miller, C. Leacock, R. Tengi, R. T. Bunker, “A semantic concordance,” in Proceedings of the workshop on Human Language Technology, HLT ’93, 303–308, Association for Computational Linguistics, Stroudsburg, PA, USA, 1993.

[26] H. Graeme, “Overcoming Linguistic Barriers to the Multilingual Semantic Web,” in P. Buitelaar, P. Cimiano, editors, Towards the Multilingual Semantic Web, 3–14, Springer Berlin Heidelberg, 2014, doi:10.1007/978-3-662-43585-4.

[27] D. Tufis, D. Cristea, S. Stamou, “BalkaNet: Aims, Methods, Results and Perspectives. A General Overview,” In: D. Tufis (ed): Special Issue on BalkaNet. Romanian JSTI, 2004.

[28] G. de Melo, G. Weikum, “Towards a universal wordnet by learning from combined evidence,” in D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, J. J. Lin, editors, CIKM, 513–522, ACM, 2009, doi:10.1145/1645953.1646020.

[29] E. Pianta, L. Bentivogli, C. Girardi, “MultiWordNet: developing an aligned multilingual database,” in Proceedings of the 1st Inter. Global Wordnet Conference, 2002.

[30] V. Nastase, M. Strube, B. Boerschinger, C. Zinn, A. Elgharfaoui, “WikiNet: A Very Large Scale Multilingual Concept Network,” in N. Caizolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias, editors, LREC, European Language Resources Association, 2010.

[31] F. Bond, R. Foster, “Linking and Extending an Open Multilingual Wordnet,” in ACL (1), 1352–1362, The Association for Computer Linguistics, 2013.

[32] M. M. Bouabdallah, N. Chabibi, N. Khaled, L. H. Belguith, F. Sadat, “Arabic WordNet semantic relations enrichment through morpho-lexical patterns,” in 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 1–6, 2013.

[33] E. Biele, “Arabic word sense disambiguation: a review,” Artif. Intell. Rev., 52(4), 2475–2532, 2019, doi:10.1007/s10462-018-9622-6.

[34] A. Budanitsky, G. Hirst, “Evaluating WordNet-based Measures of Lexical Semantic Relatedness,” Comput. Linguist., 32(1), 13–47, 2006, doi: 10.1162/coli.2006.32.1.13.

[35] N. Bouhriz, F. Benabbou, E. H. Benlahmar, Y. Zidoun, “Arabic Word Semantic similarity: a survey,” in 1st International Conference On Research In Applied Mathematics And Computer, Casablanca, Morroco, 2019.

[36] H. Rubenstein, J. W. Shavlik, editors, ICML, 296–304, Morgan Kaufmann, 1998.

[37] P. Resnik, “Using information content to evaluate semantic similarity in a taxonomy,” in IJCAI’95: Proceedings of the 14th international joint conference on Artificial intelligence, 448–453, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995.

[38] D. Lin, “An Information-Theoretic Definition of Similarity,” in J. W. Shavlik, J. W. Shavlik, editors, ICML, 296–304, Morgan Kaufmann, 1998.

[39] M. Lesk, “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone,” in SIGDOC ’86: Proceedings of the 5th annual international conference on Systems documentation, 24–26, ACM, New York, NY, USA, 1986, doi:https://doi.org/10.1145/318723.318728.

[40] Z. Zhou, Y. Wang, J. Gu, “New model of semantic similarity measuring in wordnet,” in 3rd International Conference on Intelligent System and Knowledge Engineering, volume 1, 256–261, 2008, doi:10.1109/ISKE.2008.4730937.

[41] M. A. Helou, A. Abid, “Semantic Measures based on Wordnet using Multiple Information Sources,” in KDIER 2010 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Valencia, Spain, October 25-28, 2010, 500–503, 2010.
[42] M. A. Helou, M. Palmonari, M. Jarrar, “Effectiveness of Automatic Translations for Cross-Lingual Ontology Mapping,” J. Artif. Intell. Res., 55, 165–208, 2016, doi:10.1613/jair.4789.

[43] M. A. Helou, M. Palmonari, “Cross-lingual lexical matching with word translation and local similarity optimization,” in Proceedings of the 11th International Conference on Semantic Systems, SEMANTICS 2015, Vienna, Austria, September 15-17, 97–104, 2015, doi:10.1145/2814864.2814888.

[44] Y. Regragui, L. Abouenour, F. Krieche, K. Bouzoubaa, P. Rosso, “Arabic WordNet: New Content and New Applications,” in Proceeding of the 8th Global Wordnet Conference (GWN 2016), 2016.