Building Arabic Paraphrasing Benchmark based on Transformation Rules

MARWAH ALIAN and ARAFAT AWAJAN, Princess Sumaya University for Technology, Amman, Jordan
AHMAD AL-HASAN and RAEDA AKUZHIA, Hashemite University, Zarqa, Jordan.

Measuring semantic similarity between short texts is an important task in many applications of natural language processing, such as paraphrasing identification. This process requires a benchmark of sentence pairs that are labeled by Arab linguists and considered a standard that can be used by researchers when evaluating their results. This research describes an Arabic paraphrasing benchmark to be a good standard for evaluation algorithms that are developed to measure semantic similarity for Arabic sentences to detect paraphrasing in the same language. The transformed sentences are in accordance with a set of rules for Arabic paraphrasing. These sentences are constructed from the words in the Arabic word semantic similarity dataset and from different Arabic books, educational texts, and lexicons. The proposed benchmark consists of 1,010 sentence pairs wherein each pair is tagged with scores determining semantic similarity and paraphrasing. The quality of the data is assessed using statistical analysis for the distribution of the sentences over the Arabic transformation rules and exploration through hierarchical clustering (HCL). Our exploration using HCL shows that the sentences in the proposed benchmark are grouped into 27 clusters representing different subjects. The inter-annotator agreement measures show a moderate agreement for the annotations of the graduate students and a poor reliability for the annotations of the undergraduate students.

CCS Concepts: • **Information systems → Data management systems**; **Database design and models**; **Data model extensions**;

Additional Key Words and Phrases: Paraphrasing, Arabic benchmark, transformation rules, Arabic paraphrasing benchmark, semantic similarity, inter-annotator agreement, K-means, HCL Clustering

ACM Reference format:
Marwah Alian, Arafat Awajan, Ahmad Al-Hasan, and Raeda Akuzhia. 2021. Building Arabic Paraphrasing Benchmark based on Transformation Rules. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 20, 4, Article 63 (June 2021), 17 pages.
https://doi.org/10.1145/3446770

This research is supported by the Scientific Research and Innovation Support Fund, Ministry of Higher Education, Jordan (research project ICT/2/5/2016).
Authors' addresses: M. Alian and A. Awajan, Dept. of Computer Science, Princess Sumaya University for Technology, Amman, Jordan; emails: {m.alian, awajan}@psut.edu.jo; A. Al-Hasan and R. Akuzhia, Faculty of Arts, Hashemite University, Zarqa, Jordan; emails: {Ahmadalhasan, raeda}@hu.edu.jo.

2375-4699/2021/06-ART63 $15.00
https://doi.org/10.1145/3446770

1 INTRODUCTION

Paraphrasing is a task performed to detect two texts that are a paraphrase of each other [1]. A paraphrase is a restatement of an original text to produce the same meaning in another form [2]. This task is done by using alternative words or expressions while maintaining the same meaning of the main sentence or text. For instance, the texts A and B are "paraphrases" of each other if they have the same meaning [3]. Paraphrasing and semantic similarity are useful features used to improve many natural language processing (NLP) applications, such as question answering, information retrieval, and text entailment [4].

Semantic similarity is a measure that presents the relation between words in a text according to the idea carried [5]. The measure of semantic similarity for sentences usually ranges from complete semantically equivalent to exactly unrelated in meaning. The similarity score provides a notion of intermediate similarity as the two texts may share some aspects of meaning or have semantically important differences. Moreover, the semantic similarity task in several NLP applications is considered a black box that can be evaluated independently or an internal part of the application [6].

Many approaches can be used to measure semantic similarity between sentences, which can be divided into four categories: co-occurrence-based approach, statistical corpus–based approach, feature-based approach, and word embedding–based approach. The co-occurrence-based approach represents text as bag-of-word vectors. The statistical corpus–based approach uses latent semantic analysis, which represents texts as vectors in a reduced-dimension space. The feature-based approach focuses on the similarity of words and the order between texts, aligning the words that have the same part of speech. The word net–based measure is used to compute semantic similarity between words and overlapping word orders in the two texts. Lastly, the word embedding–based approach has the ability to consider the context of the words when representing words in a distributed space [7, 26].

A benchmark is required to evaluate these tasks. Publicly available benchmarks are abundant and rare for English and Arabic short texts, respectively. These benchmarks are discussed in Section 2.

In this study, we explain the phases of building an Arabic paraphrasing benchmark in accordance with the transformation rules for Arabic language. We also explore this benchmark by using hierarchical clustering and statistical overview of the dataset over transformation rules. We obtain the sentences in the proposed benchmark by collecting sentences from Arabic books and some lexicons and generating sentences from the words in the Arabic word semantic similarity (AWSS) dataset. Finally, we use transformation rules to convert these collected sentences into pairs of sentences: the first part is the collected sentence and the second part is the transformed sentence.

To evaluate the similarity between sentence pairs, we share the proposed benchmark with a number of experts who have provided each pair with a value between 0 and 4 to represent the similarity of sentence pairs. We also asked these experts to evaluate whether the sentence pairs are paraphrased. Subsequently, we use a hierarchical clustering to explore the benchmark and group the sentences into categories. These categories represent the topic or the subject of sentences in a cluster.

The remainder of this research is organized as follows. Section 2 describes the previously constructed benchmarks. Section 3 defines the methodology used to construct the proposed benchmark. Section 4 describes the methods used for labeling the sentence pairs. Section 5 discusses the transformation rules affecting semantics. Section 6 explores the collected data. Section 7 presents the conclusion and future work.

2 RELATED WORK

Many English benchmarks for semantic similarity and paraphrasing detection have been released and used in the research community. For example, O'Shea et al. [8] construct a benchmark consisting of 65 sentence pairs labeled with human ratings for short text semantic similarity. This benchmark is widely used by the research community in evaluating methods for measuring semantic similarity between short texts. The Microsoft research paraphrase corpus [9] is an English dataset used to evaluate proposed methods for paraphrasing identification. This dataset consists of 5,800 pairs of sentences extracted from web news sites. These pairs are labeled with human annotations indicating whether each pair is paraphrased or has a semantic equivalence relationship. In the construction of this dataset, only one sentence is extracted from every news article. Also, information is provided about each sentence, such as the author and source of the sentence.

In 2006, Li et al. [25] gathered and modified 65 sentence pairs from the Collins COBUILD Dictionary and proposed a pilot short-text semantic similarity benchmark. Each pair is provided with a similarity score computed as the average of 32 human ratings [25]. For several years, the Association for Computational Linguistics has also provided considerable datasets, such as SemEval-2012, SemEval-2014, SemEval-2015, and SemEval-2017. These datasets are composed of sentences collected from corpuses and benchmarks, such as the Microsoft research paraphrase corpus and Microsoft research video description corpus. The SemEval-2015 benchmark contains all sentence pairs from tasks' benchmarks in 2012, 2013, and 2014. It consists of 8,331 pairs with a semantic similarity score that is the mean of five annotators from Amazon Mechanical Turk [24].

Only a few Arabic datasets are constructed especially for paraphrasing or semantic similarity. Researchers construct their own dataset to evaluate their approaches either for the paraphrase identification task or for semantic similarity for Arabic data. For example, Wali et al. [10] construct a dataset to evaluate their similarity measure. This dataset consists of 690 pairs of sentences provided by similarity scores. These sentences are produced from four Arabic dictionaries using dictionary definitions and examples of words. However, the dataset has not been released for the research community.

One dataset that is available online is the Arabic tweets consisting of short texts collected from tweets with positive and negative sentiments. Consequently, this dataset can be used only for sentiment analysis and not for paraphrasing identification. Another dataset for measuring similarity is the dataset released by SemEval-2017 for Semantic Textual Similarity [11]. This dataset consists of sentence pairs that have been manually translated into Arabic and labeled with a value from 0 to 5 as a similarity score. The sentence pairs are drawn from different publicly available datasets. Although the sentence pairs are not labeled for paraphrasing, we still add this label to our benchmark.

In our research, we aim to build an Arabic paraphrasing benchmark that is not collected from the web but rather is constructed by experts in accordance with the transformation rules that do not always produce identical meaning for the transformed sentence. This benchmark is available online on the site of our NLP research group[1].

3 ARABIC PARAPHRASING BENCHMARK DESIGNING

3.1 Data Collection Methodology

The first part of sentence pairs in this dataset is collected using two methods. First, we collect the sentences from the books used for teaching syntax and semantics of the Arabic language, such as AlnHw AlwADH fy qwAEd AllgAh AlErbyAh "النحو الواضح في قواعد اللغة العربية"[12], Elm Ald-

[1] http://nlp.psut.edu.jo/.

Table 1. Characteristics of Experts and Their Roles

Expert	Degree	Major	Role in the collection phase	Percentage of collected sentences
A	Associate Professor	Literature and Criticism	—Collect sentences from Arabic books —Transform sentences	45%
B	Assistant Professor	Language and Syntax	—Generate sentences from AWSS dataset —Transform sentences	55%

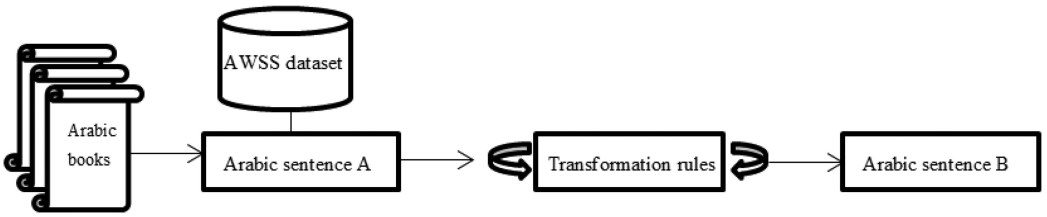

Fig. 1. Data collection process.

lAlAh "علم الدلالة" [13], Elm AldlAlAh (Elm AlmEnY) (علم المعنىعلم الدلالة) [14], AltdrybAt AllgwyAh wAlqwAEd "التدريبات اللغوية والنحوية" [15], and the Jordanian Arabic language curriculum for the fifth to ninth grades. Second, we gather the sentences generated by Arabic experts by using words from Arabic lexicons and the AWSS dataset [16]. The two experts who collect and transform the sentences in the dataset have a degree of philosophy in Arabic language and literature. They teach bachelor degree students in the Art College as their professions. The characteristics and roles of these experts are provided in Table 1.

We transform the collected and generated sentences by using the transformation rules for Arabic sentences described in [17].

To produce paraphrasing in Arabic sentences, two methods can be followed. The first is based on the following hypotheses: if two sentences are identical in all words except for the word (x) in the first sentence and the word's synonym (y) in the second sentence, then the two sentences are considered paraphrased. The second method involves the Arabic transformation rules [17]. The flow of the data collection process is shown in Figure. 1.

The sentences (Set A, the first part of sentence pairs) in the proposed benchmark have been collected from different books for teaching Arabic and constructed by the experts in accordance with the words from the AWSS dataset consisting of Arabic words with their measure of similarity. The transformation rules have been applied to these sentences (Set A) by two Arab linguists to produce the transformed sentences (Set B, the second part of sentence pairs). Both experts have a PhD in Arabic language and finalized the data collection process and the sentence transformation within three months.

However, transformation rules of the Arabic sentences do not always produce sentences that are identical in meaning. The sentences may still differ in meaning, such as " رأى موسى مازن Mousa saw Mazen" and "Mazen saw Mousa رأى مازن موسى". More details about the transformation rules are provided in the next section.

The constructed dataset consists of 1,010 sentence pairs (2,020 sentences) labeled for paraphrasing by Arabic linguists. These experts are selected from graduate students taking courses at the

Table 2. Rating Scale for Measuring Similarity

Rate	Semantic measure	
0	Unrelated in meaning	زوج الجمل لا يوجد ارتباط بينها في المعنى
1	Vaguely similar in meaning	زوج الجمل بينها تشابه ضمني في المعنى
2	Much alike in meaning	زوج الجمل التي بينها تشابه واضح اكثر من ضمني
3	Strongly related in meaning	زوج الجمل التي بينها علاقة قوية في المعنى
4	Identical in meaning	زوج الجمل المترادفة او المتطابقة في المعنى

Art College at Hashemite University. They are asked to evaluate the degree of similarity between sentences using the similarity scale provided by Alzahrani [5], as shown in Table 2.

This rating scale is provided to the graduate students, and their professors explain how they must evaluate the similarity between sentences. A value of (0) means that the sentence pairs are unrelated in meaning and are different from each other. A value of (1) means that the sentence pairs are far apart in similarity but are vaguely similar. A value of (2) is given for sentence pairs that are much alike in meaning. A value of (3) is given for sentence pairs that are strongly related in meaning. Finally, a value of (4) is given for identical sentence pairs.

3.2 Transformation Rules

The term "transformative rules" refers to Chomsky [18], the pioneer of the constructional and transformational school. However, the term "transformation" is originally constructed by Haris, and the transformation concepts and processes are described in detail in Chomsky's book of "syntactic structures" [18] and other subsequent works, such as AlKholi [17] and Benaissa [19].

The transformations that can be applied to the sentences are grouped in a set of rules referred to as "transformation rules" introduced by Chomsky [18] and described for Arabic sentences by AlKholi [17].

These rules have been limited to six patterns: permutation, deletion, addition, reduction, expansion, and replacement [17]. Permutation is done by changing the order of words, and deletion is performed by deleting an item from the sentence. Addition introduces an item to the structure of the sentence. Expansion is done by representing a word by two other words that provide the same meaning, and reduction is the replacement of two words by one word holding the same meaning [26]. More description for the transformation rules is provided in the following subsections.

Let A, B, and C be words or phrases in a sentence. Table 3 shows the transformation rules described by AlKholi [17] and our provided examples for each rule.

Transformation is a description of the relationship between the deep and surface structures of the sentence. A close link between them can be illustrated by the example in Table 4.

Among the sentences (1:a–1:d), the sentences (1:b–1:d) are mutated sentences from the sentence (1:a) in which the nouns "student الطالب" in (1:b) and "lesson الدرس" in (1:c) are transferred in a place where the nouns is the subject, with some modifications. Given that the verb "read قرأ"" is preceded nouns, a pronoun "ضمير" is left in the place previously occupied by the nouns before the transformation process. Consequently, a covert (hidden, implied) pronoun "مستتر" in (1:b) returns to the "student الطالب" and attaches (suffixes) pronoun "متصل" in (1:c) back to the "lesson الدرس" as in (1:b) and (1:c).

(1) b'. The student reads the lesson.	AlTAlb qrA [...] Aldrs	.الطالب قرأ [...] الدّرسَ
(1) c'.The lesson is read by the student.	Aldrs qrA [h] AlTAlb	.الدرس قرأ [هُ] الطالب

Table 3. Transformation Rules

Transformation rule	Represent-ation by symbols	Example	Transliteration	English translation
Permutation	A + B = B + A	تسلم الفائز الجائزة	tslm AlfA}z AljA}zAh	The winner gets the prize
		تسلم الجائزة الفائز	tslm AljA}zAh AlfA}z	The prize is given to the winner
Deletion	A + B =[...]+ B	اسألوا أهل القرية عن اللص	AsAlwA Ahl AlqryAh En AllS	Ask the villagers about the thief
	A + B =A+ [...]	اسألوا القرية عن اللص	AsAlwA AlqryAh En AllS	Ask the village about the thief
Addition	A = A + B	السماء صافية	AlsmA' SAfyAh	Clear sky
		إن السماء صافية	En AlsmA' SAfyAh	The sky is clear
Expansion	A = B + C	وددت نزول المطر	wddt nzwl AlmTr	I wanted it to rain
		وددت لو ينزل المطر	wddt lw ynzl AlmTr	I would like it to rain
Reduction	A + B = C	الجو حار بارد	Aljw HAr bArd	The weather is hot and cold
		الجو معتدل	Aljw mEtdl	The weather is fair
Replacement	A = B	شارك الأستاذ في الأمسية الأدبية	$Ark AlAstADH fy AlAmsyAh AlAdbyAh	The professor participated in the literary evening
		شارك الأستاذ في الأمسية الشعرية	$Ark AlAstADH fy AlAmsyAh Al$EryAh	The professor participated in the poetry evening

Table 4. Transformations of the Sentence "The Student Read the Lesson"

Translation to English	Transliteration	Arabic sentence
The student reads the lesson	qrA AlTAlb Aldrs	(1) a - قرأ الطالبُ الدّرسَ.
The student has read the lesson	AlTAlb qrA Aldrs	b- الطالبُ قرأ الدّرسَ.
The lesson is read by the student	Aldrs qrAh AlTAlb	c- الدرس قرأهُ الطالبُ.
The student read the lesson	qrA Aldrs AlTAlb	d- قرأ الدرسَ الطالبُ.

The sentence (1:d) is transformed from the sentence (1:a) by bringing forward the object "lesson الدرس" in the place of the subject "student الطالب.." This forwarding does not need to leave a pronoun that returns to a previous noun.

These changes and other similar changes can be called "transformations." If the sentences that differ in structure are of the same meaning, they are considered paraphrased sentences. If these changes lead to differences and variations in the meanings of the sentences, they are considered non-paraphrased sentences.

3.2.1 Permutation. Permutation is the forwarding of an element in the sentence before another element, which is known as the Arabic rule of "forward and backward."

Let A, B, and C be words or phrases in the sentence. Then, we can symbolize the permutation rule as:

$$A + B = B + A$$

Here, no word or phrase is deleted but the order of words is reversed wherein word B is forwarded and word A is moved backward in the sentence structure. For example:

(2)a. The winner received an award.

تسلّم الفائز الجائزة

فعل + فاعل + مفعول به

object + subject + verb

Thus, the sentence in the previous example is structured according to Arabic grammar. The order of its elements is as follows: the verb "الفعل" is provided first before the subject "الفاعل," and the object "المفعول به" is placed at the end. If we transfer the object element before the subject in the sentence, then we can obtain a new sentence:

(2)b. The award was received by the winner. تسلّم الجائزة الفائز

By doing this exchange, we have followed the "permutation" rule of exchange. In this case, the transformation has not led to a difference or variance in the intended meaning.

Definition 3.1. Let a sentence S be a set of word elements: $S = \{a_1, a_2, \ldots a_n\}$. Assume that the elements a_i and a_j are members of S ($a_i \in S, a_j \in S$) and ($i \neq j$). In permutation, a_i will get the position of a_j and a_j will be in the position of a_i. In other words, a_i and a_j will be replaced by each other. Then, the new sentence S' will be:

$$S' = ((S \backslash \{a_i\}) \cup \{a_j\}) \cup ((S \backslash \{a_j\}) \cup \{a_i\}).$$

3.2.2 Deletion. In this rule, the new sentence is formed by deleting one of the elements from the original structure of the sentence according to the following rule:

$$A + B = [\cdots] + B$$

A + B is converted into B, implying that the deleted element is the first element or word (A). The object "المفعول به" (أهل) in sentence (3: a) is deleted and the "added to المضاف إليه" (which is the modifier القرية) is preserved in the transformed sentence (3:b) instead of the object (the genitive phrase أهل القرية).

(3) a. Ask the villagers about that lair. اسأل أهل القرية عن ذلك الكذوب.
(3) b. Ask the village about that lair. اسأل القرية عن ذلك الكذوب.

Deletion also follows the rule: A + B = A + [· · ·]

The deleted element is B, as the deletion of the predicate "الخبر" in sentence (4:a) to be transformed into (4:b).

(4) a. The bird exists on the tree. الطائر موجود فوق الشجرة.
(4) b. The bird is on the tree. الطائر فوق الشجرة.

Definition 3.2. Let a sentence S $= \{a_1, a_2, \ldots a_n\}$. Assume that a_i is a member of S ($a_i \in S$) and it is going to be deleted from this sentence. The resultant sentence S' will be: $S' = S\{a_i\}$.

3.2.3 Addition. Addition is the increase of the structure of the sentence by adding a new element and preserving the rest of its elements as they are. The increase of auxiliaries introduces nominal sentences, such as كان-set (kāna wa-axawātuha) and إن-set (inna wa-axawātuha). This rule is described as follows:

$$A = A + B$$

In this process, the phrase or word (A) still exists in the structure of the sentence, but we add a new word or phrase (B) to the sentence structure:

(5) a. Clear sky. السماء صافية.
(5) b. The sky is clear. إنّ السماء صافية.

Definition 3.3. Let a sentence S be a set of words: S $= \{a_1, a_2, \ldots a_n\}$. Assume that a_i is going to be added as a member to this sentence. Then, the transformed sentence S' will be: $S' = S \cup \{a_i\}$.

3.2.4 Expansion. In this rule, an element is replaced by two other elements that lead to the meaning of the original element. The expansion rule is symbolized as follows:

$$A = B + C$$

In this process, A is expanded into B + C. The rule of expansion can be represented in Arabic as the conversion from explicit verbal noun "explicit participle المصدر الصريح" into verbal noun "implicit participle المصدر المؤول" as shown in sentences (6:a) and (6:b).

(6) a. I wanted it to rain. وددتُ نزولَ المطر.
(6) b. I would like it to rain. وددتُ لو ينزل المطر.

Definition 3.4. Let a sentence $S = \{a_1, a_2, \ldots a_n\}$. Assume that $A = \{a_i\}$ is a subset of S ($A \subseteq S$). In the expansion rule, the set A will be expanded to a set of two elements $A' = \{a_i, a_j\}$, where these two elements have the same meaning as a_i. Then, the resultant sentence S' will be $S' = (S \backslash A) \cup A'$.

3.2.5 Reduction. The reduction rule is the opposite of the expansion rule. Two elements are replaced by one element providing their meaning. This rule is symbolized as follows:

$$\mathrm{A} + \mathrm{B} = \mathrm{C}$$

In this process, two words or phrases (A and B) are reduced into one phrase (C) that provides a similar meaning to (A+B) together. This rule can be illustrated as shown in sentence (7: a) and its transformed version (7: b).

(7) a. The weather is hot and cold. الجو حار بارد.
(7) b. The weather is moderate. الجو معتدل.

Definition 3.5. Let a sentence $S = \{a_1, a_2, \ldots a_n\}$. Assume that $A = \{a_i, a_j\}$ is a subset of S ($A \subseteq S$). The set A will be reduced to have one element $A' = \{a_i\}$, where a_i has the same meaning as the phrase $\{a_i, a_j\}$. The resultant sentence S' will be $S' = (S \setminus A) \cup A'$.

3.2.6 Replacement. Replacement is to exchange a word with another word. This rule is symbolized as follows:

$$A = B$$

In this process, A is replaced by B. The word "literary الأدبية" in sentence (8: a) is replaced by the word "poetry الشعرية" in (8: b):

(8) a. The writer participated in the literary event. شارك الأديب في الأمسية الأدبية.
(8) b. The writer participated in the poetry event. شارك الأديب في الأمسية الشعرية.

However, we cannot judge whether the sentences are paraphrased in accordance with the transformation rule unless they have similarity in their meaning. Otherwise, the sentences are non-paraphrased.

Definition 3.6. Let S be a sentence that consists of a set of words $\mathrm{S} = \{a_1, a_2, \ldots a_n\}$, and let a_i be a member of this sentence ($a_i \in S$). Assume that b_j is going to replace member a_i in set S, where ($b_j \notin S$), ($b_j \neq a_i$), and ($i = j$). The result of removing a_i from S and adding b_j is a new sentence S':

$$S' = (S \setminus \{a_i\}) \cup \{b_j\}$$

3.3 Senses

A total of 86 sentences representing senses are included in the proposed benchmark. The sense is used in the structure of the sentence; then, different transformation rules are applied to the sentence. The sense المشترك اللفظي is considered to be specific to the vocabulary. It is like a

synonym in terms of its concern with vocabulary, not with sentences. It has an opposite definition because synonyms represent two or more words with the same meaning, whereas a sense is a word that has multiple meanings. For linguists, a sense is a word with the same letters with different meanings. The meaning of a word becomes clear when we consider its context.

For example, the word (عين eye) is used in Arabic to denote the organ of vision and the spring of water. For instance, the phrase "tears from my eyes" indicates the human eye, whereas the phrase "I drink from the eye" indicates a spring of water.

The sense in the following sentence pairs has been transformed using the transformation rules, such as addition in sentence (9), deletion in sentence (10), and replacement in sentence (11).

(9): a- شربت من عين الماء.	I drank from the spring.
b-شربت من عين الماء، فدمعت عيني.	I drank from the spring, and then my eyes got watery.
(10): a-قال لي ابني ربيع: أنت ربيع عمري.	My son Rabie said, "You are the spring of my life."
b-قال لي ابني: أنت ربيع عمري.	My son said, "You are the spring of my life."
(11): a-عثر علماء الآثار على قرن حيوان، عاش على هذه الأرض قبل مئة عام.	Archaeologists have found the horn of an animal that lived on this land a hundred years ago.
b-عثر علماء الآثار على قرن حيوان، عاش على هذه الأرض قبل قرن.	Archaeologists found an animal horn that lived on this earth a century ago.

In sentence (10: a), the word (Rabie: ربيع: rbyE) is a name of a living entity, whereas in (10: b) the word (rbyE: ربيع: spring) indicates the spring season. In sentence (11: a), the word (qrn: قرنق: horn) sense is a hard permanent outgrowth found on the heads of some animals, such as cattle. In the second part of sentence (11: b), the word (قرن: qrn) sense indicates a century.

3.4 Paraphrasing

Paraphrasing refers to sentences, whereas synonymy is related to vocabulary. For example:

(12) a. I saw the child happy. رأيت الطفل فرحاً.
(12) b. I saw the child pleased. رأيت الطفل مسروراً.

The words (happy فرحاً) and (pleased مسروراً) are synonyms because a similarity exists between the former two words caused by synonymy. Sentences (12: a) and (12: b) are judged to be paraphrased because they have similarities in meaning. Another type of paraphrasing is found in the following sentences:

(13) a. I like to eat fruit before the meal. أحب أكل الفاكهة قبل الطعام.
(13) b. I like eating apples before a meal. أحب أكل التفاح قبل الطعام.

In this example, a similarity or relatedness exists between the two words (fruits الفاكهة and apples التفاح) caused by the hyponymy relationship. The two sentences are judged as similar, and the relationship between them is entailment wherein the truth of the first sentence requires the credibility of the second one. More examples on paraphrasing are shown in Table 5.

4 TRANSFORMATION RULES AND SEMANTICS

The addition rule added a word to the structure of the sentence in a manner that indicates emphasis, definition, distinction, or negation. Semantic emphasis is provided when a word is used to confirm the meaning of a word or phrase. For example:

(14) a. I passed the president. مررت بالرئيس.
(14) b. I passed by the president himself. مررت بالرئيس نفسه.

The word "himself نفسه" is used to emphasize meeting the president "الرئيس".

Table 5. Examples of Paraphrased and Non-paraphrased Sentences

Sentence pairs		Semantic Similarity	Relationship
Khaled is the brother of Mohanad. Mohanad is the brother of Khaled.	a-1: خالد أخو مهند. b- مهند أخو خالد.	Semantically similar	Paraphrase
I do not like stingy man. I hate stingy man.	a-2: لا أحب الرجل البخيل. b- أكره الرجل البخيل.	Semantically similar	Paraphrase
I do not like a stingy man. I do not like a spiteful man	a -3: لا أحب الرجل البخيل. b- لا أحب الرجل الحقود.	Semantically not similar	Not a paraphrase
The Zoo has predators. The Zoo has animals.	a- 4: في الحديقة حيوانات مفترسة b- في الحديقة حيوانات.	Semantically similar	Entailment

Distinction is provided when a word is added to provide further specification for that word. For example, the term "golden من ذهب" is added to the word "ring خاتما" to imply that the ring is made of gold.

(15) a. The man bought a ring. اشترى الرجل خاتما.
(15) b. The man bought a golden ring. اشترى الرجل خاتما من ذهب.

Also, adding a word is used to define another word, for example, the word "reading القراءة" is added to define the word "كتاب".

(16) a. I brought a book. أحضرت كتابا.
(16) b. I brought the reading book. أحضرت كتاب القراءة.

Negative prepositions are added to sentences in order to imply negation. For example:

(17) a. The visited man honors his guest. يكرم المزور ضيفه.
(17) b. The visited man did not honor his guest. لم يكرم المزور ضيفه.

The preposition "did not لم" is added to imply the negation of the "honor يكرم" verb.

The deletion rule indicates the focus of the sentence on the idea or subject. For example:

(18) a. I went out, and it was raining. خرجت فإذا المطر يهطل.
(18) b. I went out in the rain. خرجت فإذا المطر.

When the verb "raining يهطل" is deleted from the first sentence, the focus becomes on "rain المطر". Also, in the following example:

(19) a. The thief stole the money. سرق اللص المال.
(19) b. The money was stolen. سُرق المال.

When the word "thief اللص" is deleted, the sentence becomes passive and the deletion implies that the focus is on the word "money المال".

Expansion of a word to a two-word phrase indicates further clarity and demonstration of that word or to indicate a focus on the event. For example:

(20) a. I hope the moon is rising. أتمنى طلوع القمر.
(20) b. I hope that the moon is rising. أتمنى أن يطلع القمر.

Permutation that alters the position of the word back or forward affects the meaning by identifying the owner of the case: the subject or the object. For example:

(21) a. The student reads the lesson. قرأ الطالبُ الدرس.
(21) b. The lesson is read by the student. قرأ الدرسَ الطالب.

The owner of the first sentence is the subject "student الطالب" while the owner after the permutation in the second sentence is the object "lesson الدرس".

The rule of reduction indicates briefness or conciseness. For example:

(22) a. The father and mother attended the party. حضر الأب والأم الى الحفل.
(22) b. The parents attended the party. حضر الأبوان الى الحفل.

The words "father الأب" and "mother الأم" are reduced to a single term, that is, "parents الأبوان". The second sentence provides a brief overview of who is attending the party.

The replacement rule, which replaces a word with another word, would have an effect on the meaning of the new sentence if the new word is synonymous with the replaced word. Then, the new sentence would be similar to the original sentence. However, if the new word is an antonym for the replaced word, the new sentence will be different from the original sentence. For example:

(23) a. I saw the man was happy. رأيت الرجل فرحا.
(23) b. I saw the man was sad. رأيت الرجل حزينا.

The word "sad حزينا" is an antonym for the word "happy فرحا"; thus, the new sentence has a different meaning. If the new word is "pleased مسرورا", the new sentence is the same as the original sentence.

(24) a. I saw the man was happy. رأيت الرجل فرحا.
(24) b. I saw the man was pleased. رأيت الرجل مسرورا.

5 DATASET LABELING

The sentence pairs in the proposed benchmark are labeled as similar or dissimilar by five experts of Arabic language. These experts are undergraduate and postgraduate students from the Art Faculty of Hashemite University. They are asked to evaluate the sentence pairs using the metrics previously described in Table 2. However, they differ in labeling the sentences. For example, the following sentence pair:

(25) a. What a beautiful sky. ما أحسن السماء.
(25) b. What a beautiful blue sky. ما أحسن زرقة السماء.

Two experts consider this pair similar, whereas three experts consider it dissimilar. The number of sentence pairs labeled differently by the experts with respect to the used transformation rule is shown in Table 6. The annotators provide different evaluations when sentences are transformed using the replacement rule by a percentage of 38% of the total sentence pairs, while the second percentage (32%) of the different similarity decisions is found in sentence pairs transformed using the reduction rule. This is because the use of the replacement rule may provide a different meaning to a sentence, especially when one of its elements is replaced by its antonym.

Table 6 shows that the most utilized transformation rule in constructing the Arabic paraphrasing benchmark is the replacement rule.

We label the sentence pairs in the proposed benchmark by two methods: the majority vote and the similarity score. The first method depends on the decision of the majority of experts. If three experts out of five give a label (s: similar) for a sentence pair, then the sentence pair is given a label of similar (s). The second method defines a sentence pair as similar if the similarity score is greater than or equal to 0.5. Otherwise, the pair is considered dissimilar. The similarity score is computed as the average of the rating scores of the five experts.

Table 6. Number of Pairs Labeled Differently by the Experts

Transformation Rule	Total number of pairs	Number of pairs labeled differently by the experts	Percentage of difference
Addition	188	78	0.415
Reduction	107	35	0.327
Permutation	122	9	0.074
Replacement	333	129	0.387
Deletion	129	40	0.310
Expansion	58	16	0.276
Two rules	68	33	0.485
Three rules	5	2	0.4

Table 7. K-means Experiment Results for the Two Labeling Methods

Experiment	Recall	Precision
Labeling by majority	0.875	0.791
Labeling by a similarity score $\geq$0.50	0.748	0.801

The effect of labeling sentence pairs using these methods on similarity detection is evaluated using K-means clustering [29]. The experiment is conducted by representing the sentences as vectors in the vector space by using the mean of their word weighted embedding where the Tf-Idf is provided as the weight for each word vector. K-means is applied with $k = 1{,}010$ to obtain similar sentence pairs in the same cluster. The cosine similarity is used to measure the similarity between vectors that represent sentences. The results of labeling by majority and labeling in accordance with the similarity score are represented in Table 7.

Table 7 shows that labeling sentence pairs using different methods has affected the precision and recall measures. Labeling by majority provides better recall for k-means clustering for weighted embedding than labeling using the similarity score.

5.1 Inter-annotator Agreement

For the annotations, we used two different groups. The first group consists of 52 undergraduate students, whereas the second group consists of 5 graduate students. Both groups are asked to annotate the 1,010 sentence pairs with a similarity score in the range of [0,4].

We consider the methods used to measure the inter-annotator agreement—such as Fleiss Kappa, Krippendorff's alpha, and intraclass correlation coefficient (ICC)—to obtain an idea of the reliability of the annotations.

Fleiss Kappa, which is an extension of Cohen's Kappa to work with multiple annotators and classes, is used to assess the reliability of agreement between annotators [21]. Krippendorff's alpha, which is a standard reliability measure, can be used regardless of the number of annotations, missing data, sample size, and measurement rates [22].

ICC is a reliability measure that is commonly used in inter-rater reliability analyses. It indicates the degree of correlation and agreement between annotations [23].

Table 8 compares the agreement of graduate student annotations with the agreement of undergraduate student annotations in terms of Fleiss Kappa, Krippendorff's alpha, and ICC measures.

For the graduate students, the Fleiss Kappa score is 0.298 and Krippendorff's alpha score is 0.287, indicating a fair agreement. The 95% confidence interval of the ICC estimate (0.573, 0.62) indicates moderate reliability.

Table 8. Inter-annotator Agreement Measures

Annotators	Measure	Score
Graduate Students	Fleiss Kappa	0.298
	Krippendorff's alpha	0.287
	ICC	0.6
Undergraduate Students	Fleiss Kappa	0.158
	Krippendorff's alpha	0.157
	ICC	0.391

For the undergraduate students, the Fleiss Kappa score is 0.158 and Krippendorff's alpha score is 0.157, implying poor agreement. The 95% confidence interval of the ICC estimate (0.367, 0.416) indicates poor reliability.

6 DATA EXPLORATION AND BENCHMARK EVALUATION

A normal part of the human cognitive mechanism is clustering. When a set of objects is available, humans group and arrange these objects in their minds. Unsupervised clustering algorithms have been developed in the field of machine learning to simulate this fundamental process [27].

To analyze large, multidimensional datasets, clustering techniques are commonly used [28]. Usually, data analysis tasks analyze particular data instances and their connection to other instances [27].

We perform a hierarchical clustering (HCL) to explore and analyze our dataset. HCL is an algorithm that groups similar objects into clusters wherein each cluster contains objects that are broadly similar to one another. We have to transform the sentences into vectors to apply HCL to our dataset. We use Aravec [20] to represent words as vectors in the continuous space. Each sentence is represented as the average of its content word vectors. The agglomerative hierarchical algorithm is applied with complete linkage and cosine similarity as the distance metric to determine the sentences in each cluster.

Aravec [20] is produced using the Word2Vec skip-gram model and trained on Arabic pages from the web with a vocabulary size of 145,428 and a dimension of 300 for word vectors.

A total of 27 clusters are produced in HCL. Each cluster consists of considerable vectors that are similar to one another. This similarity between sentences is measured using cosine similarity. The clusters obtained from applying HCL to our dataset are shown in Figure. 2. The results show that most of the similar sentence pairs are grouped in the same cluster, whereas dissimilar pairs are distributed in different clusters.

We further explore these clusters to find common features in their contents and come up with a description for them. For example, Cluster 25 contains the highest number of sentences (383) and 0.85 of these sentences describes events in the past. Cluster 7 gathers 231 sentences, 0.80 of which relate to training and achievement, and is the second cluster in a descending order.

The first cluster (Cluster 0) consists of 177 sentences about ethics and advice while Cluster 26 has 127 sentences representing advice, where 0.61 of the sentences in the cluster support this description. Cluster 12 collects sentences about food and tools where 0.56 of its sentences support this description, while Cluster 3 collects 101 sentences about farming, trading, and hunting, where 0.58 of those sentences approves this description.

The descriptions of other clusters, the number of sentences collected in these clusters and the percentage of their sentences that approve their description are shown in Table 9.

The total number of sentence pairs in the benchmark is 1,010 pairs. The distribution of sentence construction using transformation rules shows that the first sentence of 333 sentence pairs

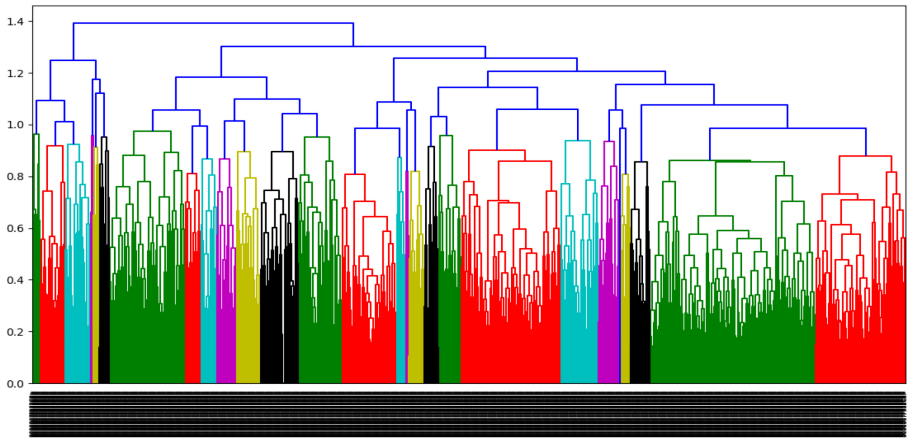

Fig. 2. Applying HCL to the Arabic paraphrasing benchmark.

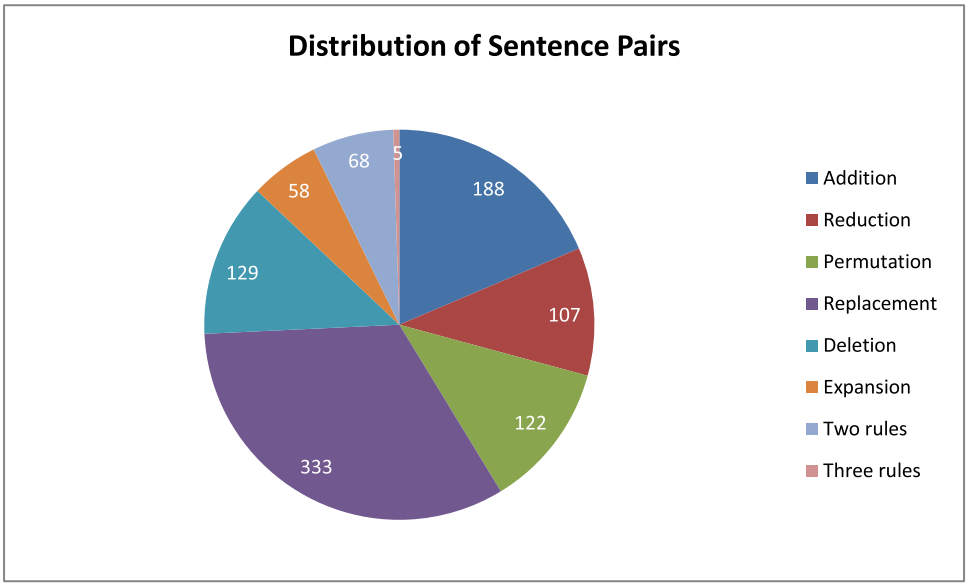

Fig. 3. Sentence pair distribution over transformation rules.

is transformed using the replacement rule in order to produce the second part of the pair, while the addition rule is used to produce the second part of 188 sentence pairs and the deletion rule is used to produce the second part of 129 sentence pairs. The permutation rule is used in 122 sentence pairs, while the reduction rule is used to transform the first part of 107 pairs and the expansion rule is used in 58 sentence pairs. The number of pairs in which two rules are used together to produce the second part of 68 sentence pairs and the second part is generated only in 5 sentence pairs using three rules together.

Figure 3 shows the distribution of sentence pairs over the transformation rules. The experts depend on the replacement and addition rules more than the others when transforming the original sentences. The least used rule in transformation is the composition of three different rules to generate a new sentence.

Table 9. Clusters Descriptions

Cluster number	Number of sentences in the cluster	Description of majority of sentences in the cluster	Percentage of sentences that approve this description
Cluster 0	177	Like/dislike, advice, and ethics	0.85
Cluster 1	7	Awards	0.86
Cluster 2	19	Cities and villages	0.58
Cluster 3	101	Hunting, trading, and farming	0.58
Cluster 4	23	Dealing with others	0.91
Cluster 5	52	Management	0.73
Cluster 6	86	Teaching and administration	0.94
Cluster 7	231	Training and achievement	0.80
Cluster 8	49	Reading, listening, painting, and buying	0.65
Cluster 9	59	Weather and nature	0.80
Cluster 10	57	Nature	0.74
Cluster 11	36	Reading and studying	0.89
Cluster 12	209	Food/drinks, animals, tools, and weather	0.56
Cluster 13	24	I am doing (personal pronoun: person speaking)	0.87
Cluster 14	45	Wars and crimes	0.80
Cluster 15	36	Patience and livelihood	0.69
Cluster 16	8	Fruits and sweets	0.50
Cluster 17	56	Workers (teachers, doctors, and soldiers)	0.70
Cluster 18	19	Emotions	0.84
Cluster 19	87	Proper noun	0.82
Cluster 20	33	Praise	0.85
Cluster 21	46	Prayer and communications	0.57
Cluster 22	13	Efforts	0.92
Cluster 23	36	Poverty and asceticism	0.75
Cluster 24	1	—	—
Cluster 25	383	Facts and events (past and future)	0.85
Cluster 26	127	Advice	0.61

After the experts label the sentence pairs as paraphrased or not, we compute the number of paraphrased pairs for each transformation rule as shown in Table 10. The highest number of paraphrased pairs is found in the replacement rule. However, the highest percentage of differently labeled pairs is provided by the permutation rule (95%). This finding means that the "forward and backward" of an element in the structure of a sentence make it ambiguous for human to determine whether the transformed sentence is paraphrased or not. The percentage of differently labeled pairs is computed by dividing the number of differently labeled pairs by the total number of pairs that have been constructed using one transformation rule.

As shown in Table 10, the evaluators differed in their assessment of 88% of the paraphrased sentences transformed by the deletion rule and differed in 81% and 79% of the sentences transformed

Table 10. Number of Paraphrased Sentence Pairs Judged Differently by the Experts

Transformation rule	Total number of paraphrased pairs	Percentage of differently labeled pairs
Addition	133	0.707
Reduction	87	0.813
Permutation	116	0.951
Replacement	219	0.658
Deletion	113	0.876
Expansion	46	0.793
Two rules	46	0.676
Three rules	5	0.833

by the reduction and expansion rules, respectively. In addition, the evaluators differed in the assessment of 71% of the paraphrased sentences that are transformed by the addition rule and 66% of the sentences that are transformed by the replacement rule.

7 CONCLUSION AND FUTURE WORK

In this research, a paraphrasing benchmark is constructed to provide the community of Arabic NLP researchers with a good standard to evaluate their work. The sentences are collected from Arabic books, and some sentences are generated by experts by using words in the AWSS dataset and from different Arabic lexicons. Six transformation rules are utilized to generate the transformed forms of the collected sentences. The benchmark is labeled by Arabic specialists with different education levels from the Art College of Hashemite University. The exploration of the benchmark shows that the sentences are grouped into 27 different clusters. The highest number of pairs are formed using the replacement rule, with 0.658 of these pairs being tagged as paraphrased. This benchmark can be considered in the future for paraphrasing identification and semantic similarity measurement tasks and other NLP applications.

ACKNOWLEDGEMENTS

We would like to thank Professor Mustafa Alian for his assistance in writing examples of the Transformation Rules and Semantics.

REFERENCES

[1] V. Vaishnavi, Madhesh Saritha, and S. Milton Rajendram. 2013. Paraphrase identification in short texts using grammar patterns. In *2013 International Conference on Recent Trends in Information Technology (ICRTIT)*. 472–477.

[2] Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. In *11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*.

[3] Peter W. Culicover. 1968. Paraphrase generation and information retrieval from stored text. *Mechanical Translation and Computational Linguistics* 11, 1 and 2 (1968), 78–88.

[4] Ngoc Phuoc An Vo, Simone Magnolini, and Octavian Popescu. 2015. Paraphrase identification and semantic similarity in Twitter with simple features. In *International Workshop on Natural Language Processing for Social Media (SocialNLP'15)*, 10–19.

[5] Salha Alzahrani. 2016. Cross-language semantic similarity of Arabic-English short phrases and sentences. *Journal of Computer Sciences* 12, 1 (2016), 1–18.

[6] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *the 9th International Workshop on Semantic Evaluation (SemEval'15)* (Denver, CO 2015). 252–263.

[7] Marwah Alian and Arafat Awajan. 2018. Semantic similarity approaches- review. In *2018 International Arab Conference on Information Technology (ACIT'18)* (Werdanye, Lebanon, 2018), 1–6.
[8] James O'Shea, Zuhair Bandar, Keeley Crockett, and David McLean. 2008. Benchmarking short text semantic similarity. *International Journal of Intelligent Information and Database Systems* 4, 2 (2008), 103–120.
[9] Bill Dolan, Chris Brockett, and Chris Quirk. 2005. *Microsoft Research Paraphrase Corpus.* (March 2005). Microsoft Research.
[10] Wafa Wali, Bilel Gargouri, and Abdelmajid Ben Hamadou. 2017. Enhancing the sentence similarity measure by semantic and syntactico-semantic knowledge. *Vietnam Journal of Computer Science* 4 (2017). 51–60.
[11] Daniel Cera, Mona Diabb, Eneko Agirrec, Iñigo Lopez-Gazpio, and Lucia Speciad. 2017. SemEval-2017 Task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. (Canada 2017). *11th International Workshop on Semantic Evaluation (SemEval-2017).*
[12] Ali AlJarem and Mustafa Ameen. 2004. *Clear grammer of Arabic language—AlnHw AlwADH fy qwAEd AllgAh AlErbyAh.* Al-Dar Almysria Alsuadia for Publishing.
[13] Ahmad Mukhtar Omar. 1998. Semantics. *Elm AldlAlAh.* Book World. Qairo.
[14] Mohammad AlKholi. 2001. Semantics. *Elm AldlAlAh (Elm AlmEnY).* Dar Al-falah. Amman.
[15] Ahmad M. Omar and others. 1999. Language and grammar exercises. *AltdrybAt AllgwyAh wAlqwAEd.* Kuwait University—Art Collage.
[16] Faaza A. Almarsoomi, James D. O'shea, Zuhair Bandar, and Keeley Crockett. 2013. AWSS: An algorithm for measuring Arabic word semantic similarity. In *2013 IEEE International Conference on Systems, Man, and Cybernetics.* 504–509.
[17] Mohammad AlKholi. 1999. Transformation rules for Arabic language. *qwAEd tHwylyAh llgAh AlErbyAh.* Dar Al-Falah. Amman.
[18] Noam Chomsky. 1957. *Syntactic Structure.* Mouton Publishers, The Hague, Paris.
[19] Abdel Haleem Benaissa. 2011. *Transfer Grammar in Arabic Phrase.* Dar Al-Kotob Al-Ilmiyah, Lebanon.
[20] Abu Bakr Soliman Mohammad, Kareem Eissa, and Samhaa R. El-Beltagy. 2017. AraVec: A set of Arabic word embedding models for use in Arabic NLP. *Procedia Computer Science* 117, (2017) 256–265.
[21] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378.
[22] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1 (2007), 77–89.
[23] Jinyuan Liu, Wan Tang, Guanqin Chen, Yin Lu, Changyong Feng, and Xin M Tu. 2016. Correlation and agreement: Overview and clarification of competing concepts and measures. *Shanghai Arch Psychiatry* 28, 2 (2016), 115–120.
[24] Adrian Sanborn and Jacek Skryzalin. 2015. *Deep learning for semantic similarity. CS224d: Deep Learning for Natural Language Processing.* Stanford, CA: Stanford University.
[25] Yuhua Li, David McLean, Zuhair Bandar, James Dominic O'Shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering* 18, 8 (2006), 1138–1150.
[26] Marwah Alian, Arafat Awajan, Ahmad Al-Hasan, and Raeda Akuzhia. 2019. Towards building Arabic paraphrasing benchmark. In *Proceedings of the 2nd International Conference on Data Science, E-Learning and Information Systems.* (2019). Article No. 17. 1–5.
[27] Joel R. Brandt, Jiayi Chong, and Sean Rosenbaum. 2006. *Interactive Clustering for Data Exploration.* Stanford University, Stanford, CA.
[28] Amir Ben-Dor, Ron Shamir, and Zohar Yakhini. 1999. Clustering gene expression patterns. *Journal of Computational Biology.* 6 (3/4). 281–297.
[29] Marwah Alian and Arafat Awajan. 2020. Factors affecting sentence similarity and paraphrasing identification. *International Journal of Speech Technology* 23, 851–859. https://doi.org/10.1007/s10772-020-09753-4

Received July 2020; revised October 2020; accepted January 2021