Towards Building Arabic Paraphrasing Benchmark

Marwah Alian
Hashemite University
Princess Sumaya University for Technology
Marwah2001@yahoo.com

Arafat Awajan
Princess Sumaya University for Technology
Amman, Jordan
Awajan@psut.edu.jo

Ahmad Al-Hasan
Hashemite University
Zarqa,Jordan
Ahmadalhasan@gmail.com

Raeda Akuzhia
Hashemite University
Zarqa,Jordan
Akuzhiar@gmail.com

ABSTRACT
This research describes a paraphrasing benchmark of Arabic sentences for evaluating algorithms developed to measure Semantic Similarity for sentences and Paraphrasing identification task. The sentences are formed based on a set of rules for Arabic paraphrasing. These sentences are constructed from words of different types of Arabic books; educational, semantic science, lexicons and AWSS dataset. The benchmark consists of 1011 sentence pairs and labeled with human ratings for semantic similarity between sentence pair.

CCS Concepts
• Information systems→Database management system engines • Datasets→Information Technology.

Keywords
Paraphrasing; Arabic dataset; transformation rules; Arabic paraphrasing dataset.

1. INTRODUCTION
Paraphrasing is the task of detecting if two texts are a paraphrase of each other [1] in other words paraphrasing is a restatement of a text in order to produce the same meaning in another form [2]. This task together with semantic similarity has shown to be useful features for improving many NLP applications such as Question Answering, Information Retrieval, text entailment and others [3].

However, semantic similarity is defined as a measure that presents the relation between words in a text according to the idea carried [4]. The measure of semantic similarity for sentences usually has the range from complete semantically equivalence to exactly unrelated in meaning, and the similarity score gives 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 as a black box that can be evaluated independently or as an internal part of the application [5].

In order to evaluate these tasks a benchmark is required. Benchmarks for English short text are available and used in several researches. For example O'Shea et al. [6] construct a benchmark for short text semantic similarity that consists of 65 sentence pairs labeled with human ratings for similarity.

DATA'19, December 2–5, 2019, Dubai, United Arab Emirates

ACM ISBN 978-1-4503-7284-8/19/12…$15.00
https://doi.org/10.1145/3368691.3368708

This benchmark is widely used by research community in evaluating methods for measuring semantic similarity between short texts. Also, Microsoft Research Paraphrase Corpus [7] is an English dataset for evaluating proposed methods for paraphrasing identification. This dataset consists of 5800 pairs of sentences that have been extracted from web news then these pairs are labeled with human annotations that indicates if each pair are paraphrased or if it has semantic equivalence relationship. They extract only one sentence from every news article that they collect. Also, they provide information about each sentence such as the author and the source of the sentence.

On the other hand, No Arabic dataset is constructed for paraphrasing or semantic similarity. Researchers construct their own dataset for evaluating their approaches either for paraphrase identification task or for semantic similarity for Arabic data. Many of these datasets are not released for the research community but some Arabic dataset are available online such as Arabic tweets dataset which consists of short texts collected from tweets and provided with positive and negative sentiments which could be only used in sentiment analysis researches.

This paper is organized as follows; Section 2 describes the collection methodology for the data. In Section 3 an explanation for the transformation rules of Arabic sentences is introduced while in Section 4 a definition for paraphrasing is given. Finally, a conclusion is provided in Section 5.

2. DATA COLLECTION METHODOLOGY
In this dataset, some Arabic sentences are collected from books used for teaching syntax and semantic of Arabic language such as AlnHw AlwADH fy qwAEd AllgAh AlErbyAh " النحو الواضح في قواعد اللغة العربية" [8], Elm AldlAlAh "علم الدلالة"[9], Elm AldlAlAh (Elm AlmEnY) علم الدلالة (علم المعنى) [10], AltdrybAt AllgwyAh wAlqwAEd "التدريبات اللغوية والنحوية" [11] and Jordanian Arabic language curriculum for the fifth to ninth grades. Other sentences are created by Arabic Experts using words from Arabic lexicons and Arabic Word Semantic Similarity (AWSS) dataset [12].

Each sentence is transformed into another sentence using transformation rules for Arabic sentences [13]. In order to produce paraphrasing in Arabic sentences, two methods can be followed; the first one is based on the hypotheses: if two sentences are identical in all words except the word (x) in the first sentence and its synonym (y) in the second sentence, then the two sentences are considered paraphrased. While the second one is based on Arabic transformation rules [13]. The flow of data collection process is shown in Figure 1.

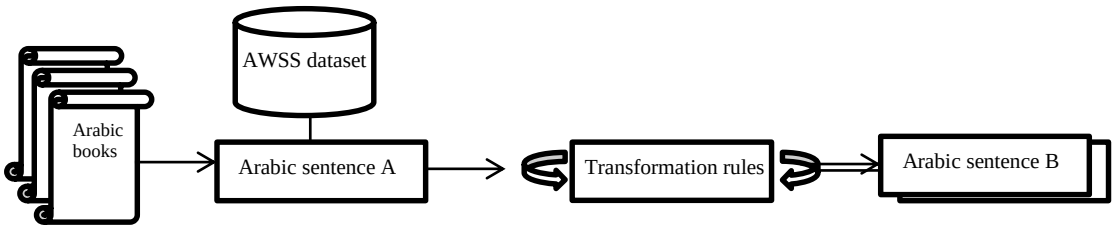

Figure 1: Data collection process

The transformation rules have been applied to the collected sentences (set A) by two experts in Arabic language to produce the transformed sentences (set B). Both experts have the degree of philosophy in Arabic and they took three months in the process of collecting and transforming sentences.

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

The constructed dataset consists of 1011 sentence pairs (2022 sentences) labeled for paraphrasing by experts of Arabic. The experts are selected from different levels ranged from undergraduate to graduate students. Then they were asked to evaluate the degree of similarity between sentences using the similarity scale provided by Alzahrani [4] as shown in Table 1.

Table 1. 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	very 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

3. TRANSFORMATION RULES

The term "transformative rules" refers to Chomsky [14], the pioneer of the constructional and transformational school; however, the term "transformation" was originally constructed by Haris, while transformation concepts and processes were described in details in Chomsky's book "syntactic structures" [14] and other subsequent works such as Akholi [13] and Benaissa [15].

The transformations that can be applied to the sentences are grouped in a set of rules referred to as “transformation rules” introduced by Chomsky [14] and described for Arabic sentences by Alkholi [13]

These rules have been limited to the following patterns: permutation, deletion, addition, reduction, expansion, and replacement. Permutation is done by changing the order of words while deletion is performed by deleting an item from the sentence which is the inverse of the addition rule which add an item to the structure of the sentence. The expansion is done by representing a word by two other words that provides the same meaning while reduction is the replacement of two words by one with the same meaning. More description for the transformation rules is given in the following subsections.

Let A, B, and C be words or phrases in the sentence. Then, the transformation rules can be symbolized in Table 2 as described by Al-kholi [13] and we provide each rule with an example.

Table 2. Transformation rules

Transformation rule	Representation by symbols	Example	Translitration
Permutation	A+B = B+A	تسلم الفائز الجائزة تسلم الجائزة الفائز	tslm AlfA}z AljA}zAh tslm AljA}zAh AlfA}z
Deletion	A + B = […] + B A + B = A + […]	اسألوا أهل القرية عن اللص اسألوا القرية عن اللص	AsAlwA Ahl AlqryAh En AllS AsAlwA AlqryAh En AllS
Addition	A=A+B	السماء صافية إن السماء صافية	AlsmA' SAfyAh En AlsmA' SAfyAh
Expansion	A = B+C	وددت نزول المطر وددت لو ينزل المطر	wddt nzwl AlmTr wddt lw ynzl AlmTr
Reduction	A + B = C	الجو حار بارد الجو معتدل	Aljw HAr bArd Aljw mEtdl
Replacement	A = B	شارك الأستاذ في الأمسية الأدبية شارك الأستاذ في الأمسية الشعرية	$Ark AlAstADH fy AlAmsyAh AlAdbyAh $Ark AlAstADH fy AlAmsyAh Al$EryAh

The transformation is a description of the relationship between the deep structure of the sentence and the surface structure of the sentence; a close link between them can be illustrated by the example in Table 3.

Table 3. Transformations of the sentence "The student read the lesson"

Translation to English	Translitration	Arabic sentence
The student read 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.

In the sentences (1: a - 1: d), the sentences (1: b - 1: d) are mutated sentences from the sentence (1 :a), in which the name "studentالطالب " in (1: b) and the name "lessonالدرس " in (1: c) are transferred in a place where the name is the subject, with some modifications; since preceding the verb "readقرأ " by a name leaves a pronounضمير in the place previously occupied by the name before the transformation process, that will be a Covert (hidden, implied) pronoun "مستتر" in (1 : b) which returned to the "student الطالب" and attached (suffixes) pronoun "متصل" in (1: c) back to the "lesson الدرس " as in 1: b', 1: c')

(1): b'- الطالب قرأ [...] الدّرسَ. Aldrs [...] AlTAlb qrA

(2): c'- الدرس قرأ [هُ] الطالب. Aldrs qrA [h] AlTAlb

While the sentence (1: d) was transformed from the sentence (1: a) by bringing forward the object "lesson"الدرس in the place of the subject "student"الطالب , and this forwarding does not need to leave a pronoun that returns to a previous name.

These changes and other similar changes can be called "transformations" if the sentences that differ in their structure are of the same meaning, and then we can judge them as paraphrased sentences; if these changes lead to differences and variations in the meanings of the sentences, they are judged as non-paraphrased sentences.

3.1 Permutation

Permutation is the forwarding of an element in the sentence on another element, which is known as 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. For example:

(2): a- تسلّم الفائز الجائزة The winner receives the award

فعل + فاعل + مفعول به

object + subject+ verb

So, the sentence in the previous example is structured according to Arabic grammar as it can be seen in the order of its elements where the verb "الفعل" is provided first, then the subject "الفاعل" and finally the object "المفعول به", if we forward the object element to be before the subject in the sentence, then we will get the new sentence:

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

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.

3.2 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 = […] + B

In this process A+B is converted into B only which means that the deleted element is the first element A, as the object "المفعول به" is deleted in sentence (3:a) and keeping the 'added to المضاف إليه '(which is the modifier) and replace it instead of the object (the genitive phrase أهل القرية) as in the transformed sentence (3:b)

(3):a-اسأل أهل القرية عن ذلك الكذوب Ask the villagers about that lair

(3): b- اسأل القرية عن ذلك الكذوب. Ask the village about that lair

Also, deletion could follow the rule:

A + B = A + […]

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

(4): a- الطائر موجود فوق الشجرة. The bird exists above the tree

(4): b- الطائر فوق الشجرة. The bird is above the tree.

3.3 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. As in the increase of Auxiliaries introducing nominal sentences such as كان set (kāna wa-axawātuha) and إن set (inna wa-axawātuha) and others. This can be represented by the rule:

A=A+B

In this process the phrase or word A still exist in the sentence but we add a new word or phrase (B).

As in the following example:

(5): a- السماء صافية. The clear sky

(5): b- إنّ السماء صافية. The sky is clear.

3.4 Expansion

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

A=B+C

In this process A is expanded into B+C. The rule of expansion can be represented in Arabic as what happen in converting from explicit verbal noun "Explicit participle المصدر الصريح" into converted verbal noun "Implicit participle المصدر المؤول" as in sentence (6:a) and (6:b).

(6): a- وددتُ نزولَ المطر. I wish the rain fall

(6): b - وددتُ لو ينزل المطر. I wish the rain would fall

3.5 Reduction

What happens in the reduction rule is the opposite of what is done in the expansion rule, where two elements are replaced by one element that will give their meaning, the reduction rule is symbolized by:

A + B = C

In this process two words or phrases (A and B) are reduced into one phrase (C). This can be illustrated in Arabic as in sentence (7:a) and its transformed version (7:b).

(7): a- الجو حار بارد. The weather is hot and cold

(7): b- الجو معتدل. The weather is moderate

3.6 Replacement

The replacement is to exchange a symbol with another symbol which can be symbolized by:

A=B

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

(8): a- شارك الأديب في الأمسية الأدبية
The writer participated in the literary event

(8): b- شارك الأديب في الأمسية الشعرية.
The writer participated in the literary event

However, it cannot be judged that the sentences are paraphrased according to the transformation rules unless they have similarity in their meaning, otherwise the sentences are non-paraphrased.

4. PARAPHRASING

Paraphrasing refers to sentences, while synonymy is related to vocabulary, for example:

(9):a- رأيت الطفل فرحا. I saw the child happy

(9):b- رأيت الطفل مسرورا. I saw the child pleased

The relationship between (happy فرحا) and (pleased مسرورا) are synonyms since there is a similarity between the former two words caused by synonymy while sentences (9) and 2 are judged to be paraphrased, as they have similarities in meaning. For example:

(10):a- أحب أكل الفاكهة قبل الطعام. I like to eat fruit before food

(10):b- أحب أكل التفاح قبل الطعام. I like eating apples before food

In this example, there is a similarity or relatedness between the two words (fruits الفاكهة and apples التفاح) caused by the hyponymy relationship. While the two sentences are judged as similar and the relationship between them is entailment where the truth of the first sentence requires the credibility of the second one. More examples on paraphrasing are shown in Table 4.

Table 4. Examples of paraphrased and non-paraphrased sentences.

Paraphrased sentences (semantically similar)	Non paraphrased sentences (semantically not similar)
a-1 : خالد أخو مهند. khaled is the brother of Mohanad b- مهند أخو خالد. Mohanad is the brother of Khaled	a -1: لا أحب الرجل البخيل. I don't like stingy man b- لا أحب الرجل الحقود. I don't like spiteful man
a-2: لا أحب الرجل البخيل. I don't like stingy man ب- أكره الرجل البخيل. I hate stingy man	a -2: في الحديقة حيوانات مفترسة The Zoo has predators b- في الحديقة حيوانات The Zoo has animals

5. CONCLUSION

In this research, a paraphrasing benchmark is constructed to provide the community of Arabic NLP researchers with a gold standard to evaluate their work. The sentences are collected from Arabic books and some sentences are generated by experts using words form Arabic lexicons and AWSS dataset. Six transformation rules are utilized to generate the transformed form of the collected sentences. The dataset is labeled by Arabic specialists with different levels from Art College from Hashemite University. For future work we are conducting a statistical analysis for this dataset to study the factors that may affect human rating.

6. REFERENCES

[1] Vaishnavi V, Saritha M, Milton R S. (2013). Paraphrase Identification in Short Texts using Grammar Patterns. In *2013 International Conference on Recent Trends in Information Technology (ICRTIT)*, 472-477.

[2] Fernando S., Stevenson M. (2008). A semantic similarity approach to paraphrase detection. In *the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*.

[3] Vo, N. P. A., Magnolini, S., Popescu, O. (2015). Paraphrase Identification and Semantic Similarity in Twitter with Simple Features. In *International Workshop on Natural Language Processing for Social Media (SocialNLP 2015)*, 10-19.

[4] Alzahrani, S. (2016). Cross-Language Semantic Similarity of Arabic-English Short Phrases and Sentences. *Journal of Computer Sciences*, 12, 1, 1-18.

[5] Agirrea, E., Baneab, C., Cardiec, C., Cerd, D. , Diabe, M., Gonzalez-Agirrea, A., Guof, W., Lopez-Gazpio, I., Maritxalara, M., Mihalceab, R., Rigaua, G., Uriaa, L., Wiebeg, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *the 9th International Workshop on Semantic Evaluation (SemEval 2015)* .Denver, Colorado, 252–263.

[6] O'Shea J., Bandar Z., Crockett K., McLean D. (2008). *Pilot Short Text Semantic Similarity Benchmark Data Set: Full Listing and Description*.

[7] Dolan, B., Brockett, C., and Quirk, C. (2005). *Microsoft Research Paraphrase Corpus*. Microsoft Research, March 2.

[8] على الجارم، مصطفى أمين, AlJarem, A., Ameen, M. *النحو الواضح في قواعد اللغة العربية AlnHw AlwADH fy qwAEd AllgAh AlErbyAh*. الدار المصرية السعودية للطباعة والنشر, 2004.

[9] Umar, A. M. (1998) .عمر, احمد مختار *علم الدلالة Elm AldlAlAh*. عالم الكتب, القاهرة.

[10] Alkholi, M A. (2001). الخولي, د. محمد علي *(علم المعنى) Elm AldlAlAh (Elm AlmEnY)*. دار الفلاح, عمّان.

[11] Umar, A. M. (1999). وآخرون, أحمد مختارعمر. *التدريبات اللغوية والقواعد النحوية AltdrybAt AllgwyAh wAlqwAEd*. جامعة الكويت - كلية الآداب , 1420.

[12] Almarsoomi, F. A., O'shea, J. D., Bandar, Z., & Crockett, K. (2013). AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 504-509.

[13] AlKholi, M. (1999). الخولي, محمد. *قواعد تحويلية للغة العربية qwAEd tHwylyAh llgAh AlErbyAh*. دار الفلاح للنشر والتوزيع, عمان.

[14] Chomsky, N. (1957). *syntactic structure*. Mouton publishers, the Hague. Paris.

[15] Benaissa, A. (2011). بن عيسى, عبدالحليم. *Transfer grammar in Arabic Phrase, القواعد التحويلية في الجملة العربية*. دار الكتب العلميةDar Al-Kotob Al-Ilmiyah, لبنان.