Constructing and Augmenting a Bidirectional Paraphrases Dataset from an English-Arabic Subtitling Parallel Corpus

Fahad AlGhamdi [1], Abdelati Hawwari [2], Mohamed Attia [3, *]

[1] Assistant professor at the Department of Computer Science, Al-Baha University, Al-Baha - Saudi Arabia, fghamdi@bu.edu.sa

[2] Datalex4ai, Santa Clara – California - USA, abdelati@datalex4ai.com

[3] Senior Consultant at the Engineering Company for the Development of Digital Systems (RDI), Giza –Egypt
[*] Corresponding author: tel. 001-647-530-1235, e-mail: Mohamed.Attia.NLP@gmail.com

Abstract: ***Paraphrasing is one of the major yet the most challenging tasks of the deep semantic analysis of natural languages. In this paper we present a novel algorithm that operates on a big parallel text corpus and automatically generates the paraphrases of the two natural languages of the corpus. Like several previously crafted algorithms in this regard, our algorithm exploits the bidirectional translation provided by the big parallel text corpora to infer couples of synonymous phrases, however, our algorithm is simpler and more efficient. Moreover, our algorithm is the only one that constructs the whole paraphrase through its run without any need for further post processing. We implemented and ran our algorithm on the English-Arabic text corpora from the 2018 version of the OpenSubtitles (OPUS) parallel text corpora, and through the statistical evaluation of random samples we found that the semantic quality among the phrases of the automatically generated paraphrases to be interestingly superb.***

Keywords: ***bidirectional semantic augmentation; parallel corpora; paraphrase; paraphrasing; phrase; semantic analysis; synonymous***

1 INTRODUCTION

In a natural language, a *paraphrase* is a set of the different phrases that expresses the same (or almost the same) meaning or communicative function [16]; i.e. a paraphrase is the phrase-level analogue of a *synset* [1]. Having the set of the most common paraphrases of a natural language is invaluable for several NLP applications.

Information retrieval, for example, becomes more effective when it considers paraphrases for flexibly comparing phrases in the queries versus the phrases in the pools of (*indexed*) text corpora. The performance of plagiarism detection tools also gets boosted when paraphrases are invoked while comparing the manuscript at question versus other candidate manuscripts. Similarly, generative language models - especially; generative Large Language Models (LLMs) - gets more flexibly effective when they make use of the paraphrases of the natural language being treated. Paraphrases also enrich classic as well as Machine-Assisted Language Learning (MALL). This is just to mention a few examples of the potential benefits of paraphrases in the field of natural language processing.

In this paper, we present an attempt to automatically extract the most common paraphrases from large parallel English-Arabic text corpora. The extracted paraphrases are bidirectional in the sense that each extracted Arabic paraphrase is accompanied with its corresponding English paraphrase, and vice versa. In the next section after this introduction, we describe the structure of the bidirectional paraphrases dataset produced by this attempt.

We developed and deployed a novel algorithm that we call *Bidirectional Semantic Augmentation* for automatically constructing this bidirectional paraphrases dataset. Our novel algorithm along with the whole process of extracting the bidirectional paraphrases from the raw drama subtitling parallel corpora is described in the third section of this paper.

The raw parallel text corpora we start with is part of OpenSubtitles2018 (OPUS: *https://opus.nlpl.eu/OpenSubtitles/corpus/version/OpenSubtitles*) which is a large dataset of translated movie

[1] Set of *synonymous words* in a Word Net.

subtitles in 60 languages, including Arabic and English. It contains 3.7 million movie and TV subtitles, which are broken down into 3.4 billion sentences and 22.2 billion words. [9] [15] While the total number of Arabic tokens in the 2018 version of OPUS is 159,043,254, the total number of English tokens is 197,098,481, and the total number of pairs of sentences is 29,823,188.

We selected this raw parallel English-Arabic text corpora as it abundantly represents a vibrant everyday conversational language that both covers a wide range of genres and the contemporary era extending from 1930's to 2010's. The corpora's informal and colloquial language, as seen in subtitles, makes it a precious resource for computer-assisted language learning applications. However, the downside of this selection is the relatively poor quality of this kind of parallel corpora due to the short time-to-market as well as the tight budget that are both typical in the subtitling industry; esp. in light of a combination of a limited supply of skilled translators and a high demand of subtitled foreign English-based media for a plethora of cinemas and TV channels in the Arab world.

While the fourth section concludes this paper by presenting a random-sample based evaluation of the semantic quality among the phrases of the bidirectional paraphrases dataset produced by our first attempt, it should be noted that the tools set we built for automatically constructing bidirectional paraphrases from parallel corpora can also be deployed to work on different raw parallel text corpora with different qualities and language pairs in order to advance the state-of-the-art in building this type of LRs that represent natural language at a deep semantic level.

2 STRUCTURE OF THE BIDIRECTIONAL PARAPHRASES DATASET

The bidirectional paraphrases dataset produced by our work consists of a set of bidirectional paraphrase entries; each entry consists in turn of an English paraphrase structure and its corresponding Arabic paraphrase structure. A paraphrase structure is essentially a list of synonymous phrase cells; each of them contains the string of the phrase along with the phrase occurrence in the parallel corpora. This list is put in a descending order according to its phrase occurrences, and the most frequent phrase (on top of the list) is called the *head phrase* in the paraphrase. Figures 1, 2 and 3 below illustrate three sample examples out of the 181,467 bidirectional paraphrases produced by our work. [2]

English phrases of the paraphrase		Arabic phrases of the paraphrase	
Number of synonymous English phrases = 11 Sum of the frequencies of English phrases = 50		**Number of synonymous Arabic phrases = 5 Sum of the frequencies of Arabic phrases = 28**	
You're a liar	10		
You're a liar !	7		
You liar	6		
You're lying	5	10	أنت كاذب
You are a liar	5	8	أنت كاذبة
you're a liar	5	5	أنت كذاب
You liar !	3	3	انتي كاذبة
You are such a liar	3	2	إنك كاذبة
You are a liar !	2		
You're full of shit	2		
You're lying !	2		

Figure 1: A first illustrative example of a bidirectional English-Arabic paraphrase

[2] We mean that the *whole* English part of a paraphrase entry corresponds to the *whole* Arabic part of the same paraphrase entry, and vice versa.

English phrases of the paraphrase		Arabic phrases of the paraphrase	
Number of synonymous English phrases = 12 Sum of the frequencies of English phrases = 96		**Number of synonymous Arabic phrases = 15 Sum of the frequencies of Arabic phrases = 91**	
		31	**أتوسل إليك**
I beg you	31	12	**أرجوك**
I'm begging you	14	8	**اتوسل إليك**
I beg of you	13	6	**أنا أتوسل إليك**
I beg you !	7	6	**أتوسل إليكم**
I am begging you	6	5	**أترجاك**
Please	6	4	**أستجديك**
I'm begging you !	5	3	**من فضلك**
I implore you	4	3	**أنا أرجوك**
I beseech you	3	3	**أنا أتوسل إليكم**
I beg of you !	3	2	**أرجوك !**
Please !	2	2	**أرجوكم**
please	2	2	**أنا أترجاك**
		2	**أنا اتوسل إليك**
		2	**أستجداك**

Figure 2: A second illustrative example of a bidirectional English-Arabic paraphrase

English phrases of the paraphrase		Arabic phrases of the paraphrase	
Number of synonymous English phrases = 14 Sum of the frequencies of English phrases = 65		**Number of synonymous Arabic phrases = 14 Sum of the frequencies of Arabic phrases = 42**	
It doesn't make sense	10	10	**هذا غير منطقي**
That doesn't make any sense	8	5	**هذا غير مفهوم**
This doesn't make any sense	7	5	**هذا لا يعقل**
That doesn't make sense	6	2	**هذا يبدو غير مفهوم**
Doesn't make any sense	6	2	**ذلك لا يعقل**
This doesn't make sense	5	2	**ليس منطقيا**
It makes no sense	4	2	**لا يبدو منطقيا**
It doesn't make any sense	4	2	**لا معنى له**
that doesn't make any sense	3	2	**هو لا يصبح مفهوما**
That makes no sense	3	2	**لا يبدو هذا منطقيا**
That doesn't even make sense	3	2	**ليس له معنى**
Makes no sense	2	2	**الأمر غير منطقي**
that doesn't make sense	2	2	**لا معنى لهذا**
It's illogical	2	2	**لا يبدو ذلك منطقيا**

Figure 3: A third illustrative example of a bidirectional English-Arabic paraphrase

Our bidirectional paraphrases dataset is automatically generated through our *bidirectional semantic augmentation* algorithm - described in the next section of this paper - from a bidirectional phrase thesaurus that is in turn automatically generated from the raw parallel corpora. Therefore, it is necessary to describe this bidirectional phrase thesaurus in the rest of this section.

This bidirectional thesaurus is composed of two main halves; a phrase-level English-to-Arabic thesaurus and a phrase-level Arabic-to-English thesaurus, each of them consists of a list of lexical entries.

The head component in each English-to-Arabic lexical entry is an English phrase from the subtitling corpora, along with its corresponding Arabic phrases (i.e. possible Arabic translations) that occur in the subtitling corpora. Each of these lexical entries registers also the number of the occurrences of the English phrase in the subtitling corpora, as well as the number of occurrences of each of its corresponding Arabic phrases in the entry.

Similarly, the head component in each Arabic-to-English lexical entry is an Arabic phrase from the subtitling corpora, along with its corresponding English phrases (i.e. possible English translations) that occur in the subtitling corpora. Each of these lexical entries registers also the number of the occurrences of the Arabic phrase in the subtitling corpora, as well as the number of occurrences of each of its corresponding English phrases in the entry.

In order to be a head component of a lexical entry, the key phrase *must at least occur twice* in the subtitling corpora. In order to stay among the corresponding phrases of the key phrase of a given lexical entry, the phrase *must at least occur twice* within the same lexical entry. If all the corresponding phrases of a lexical entry occur only once within the same entry, the whole lexical entry is omitted from our thesaurus.

Therefore, our thesaurus registers a possible correspondence between a source phrase and a target phrase if and only if it is statistically reliable. Our criterion for statistical reliability dictates an occurrence ≥ 2, which relies on the well-known fact that simple as well as compound natural language units in large corpora are statistically distributed so that the frequencies r of language units versus the frequencies of frequencies of language units $f(r)$ tend to be related via an inverse power law of the form $f(r) = a \cdot r^{-b}; b \geq 1$, that is known as *Zipf-Mandelbrot* probability distribution. [10] In our case such a distribution means that so many (actually *most* of the) phrases occur *only once* in the raw subtitling corpora. Such phrases are called *singletons*, and may be well regarded as statistical noise that should be eliminated from a lexicon concerned with *common* phrase translations like ours.

While figure 4 illustrates two examples of our English-to-Arabic lexical entries, figure 5 illustrates two examples of our Arabic-to-English lexical entries.

	English key phrase	Occurrences
	Good day	183
Serial	**Arabic target phrases**	**Occurrences**
01	يوم جيد	33
02	يوم سعيد	25
03	طاب يومك	18
04	مرحبا	10
05	طاب يومكم	10
↓ :	↓ :	↓ :
23	يوما جيدا	2
24	أتمنى لك يوما طيبا	2
25	مساء الخير	2

	English key phrase	Occurrences
	Good morning	409
Serial	**Arabic target phrases**	**Occurrences**
01	صباح الخير	110
02	عمت صباحا	55
03	صباح الخير !	50
04	مرحبا	50
05	طاب صباحك	10
↓ :	↓ :	↓ :
30	صباح الخير يا عزيزي	2
31	عمتم صباحا	2
32	صباح الخير جميعا	2

Figure 4: Two examples of English–to-Arabic lexical entries in the thesaurus

	Arabic key phrase	Occurrences
	طاب صباحك	37
Serial	**English target phrases**	**Occurrences**
01	Morning	15
02	Good morning	10
03	good morning	7
04	Morning to you	5

	Arabic key phrase	Occurrences
	نهارك سعيد	23
Serial	**English target phrases**	**Occurrences**
01	Good day	7
02	Have a nice day	4
03	Hello	4
04	Good morning	3
05	Good day to you	3
06	Have a good morning	2

Figure 5: Two examples of Arabic-to-English lexical entry in the thesaurus

Based on the aforementioned statistical criterion, our thesaurus has 147,852 English-to-Arabic lexical entries, and the sum of the occurrences (in the subtitling parallel corpora) of all these English *key source* phrases in these entries is 2,163,192 English phrases. The number of all the Arabic *target* phrases corresponding to all the English *key* phrases is 1,842,106 with an average of 12.46 target Arabic phrases per each key source English phrase. Moreover, the sum of the occurrences (in the subtitling parallel corpora) of all the target Arabic phrases in all the lexical entries is 8,127,665.

On the other hand, our thesaurus has 181,467 Arabic-to-English lexical entries, and the sum of the occurrences (in the subtitling parallel corpora) of all these Arabic *key source* phrases in these entries is 1,868,649. The number of all the English *target* phrases corresponding to all the Arabic *key source* phrases is 320,565 with an average of 1.77 distinct target English phrases per each key source Arabic phrase. Moreover, the sum of the occurrences (in the subtitling parallel corpora) of all the target English phrases is 1,040,387.

3 THE BIDIRECTIONAL SEMANTIC AUGMENTATION ALGORITHM

A. A review on previous work

Before unfolding the details of our new algorithm for extracting bidirectional/bilingual paraphrases from parallel text corpora, we start with a brief review of the most significant previous R&D works on automatic paraphrasing from parallel corpora over the past two decades.

Dolan, Quirk, and Brockett [6] along with Dolan and Brockett [5] made a major contribution in building paraphrases dataset. Their efforts have been crowned by building the Microsoft Research Paraphrase Corpus (MSRPC) that is broadly used in NLP research works. They capitalized on the repetitive occurrences of phrases in news articles collected from the World Wide Web.

Quirk, Brockett, and Dolan [12] proposed *monolingual machine translation* to generate paraphrases. This method treats paraphrasing as a translation problem where phrases/sentences are translated into another form within the same language using bilingual machine translation.

Denkowski and Lavie [4] produced METEOR-NEXT and paraphrase tables in METEOR to improve evaluation support for multiple languages. Their techniques tune paraphrase lists with machine translation metrics, which provides improved evaluation accuracy.

Tschirsich and Hintz [14] used crowd-sourcing to recognize paraphrases, which proved the effectiveness of human intelligence at recognizing paraphrases. This approach showed how human translators and automatic systems can work together to build high quality paraphrase datasets.

Ganitkevitch, Van Durme, and Callison-Burch [8] along with Ganitkevitch and Callison-Burch [7] presented the paraphrases database (PPDB) and its multilingual extension. PPDB extracts paraphrases from multiple bilingual parallel corpora, which affords a comprehensive resource to generate paraphrases across multiple languages.

Pavlick et al. [11] expanded PPDB to PPDB 2.0 with better paraphrase ranking, and also included fine-grained entailment relations, word-embedding similarities, and style annotations.

Mathias Creutz [3] built *Opusparcus*; a paraphrase corpus for six European Languages from the big volume of multilingual parallel dialogues from the subtitling of drama and TV shows available in OpenSubtitles2016, which affords a rich dataset for training and evaluating the ranking models of paraphrased pairs of phrases/sentences.
Then comes our algorithm; *Bidirectional Semantic Augmentation*, that we unfold later in this section of the paper after giving a hint on the necessary preprocessing of the raw parallel corpus before being fed to this algorithm.

B. Preprocessing the raw parallel corpora

Segmentation, data cleaning, and spell-checking are the three main tasks of preprocessing the raw parallel corpus. Segmentation starts with dividing dialogue sentences into simple sentences; relying on the punctuators marking the end-of-sentence namely {*full stop*, *exclamation mark*, *question mark*}. We excluded sentence pairs with no correspondence between their inner punctuators namely {*comma*, *semicolon*, *colon*}. We only accept the sentence pairs whose punctuators are matching in number and types. Then comes the segmentation into phrases; relying on the punctuators marking the end-of-phrases namely {*comma*, *semicolon*, *colon*}, the parallelization between each pair of English-Arabic phrases, and the elimination of problematic phrases where the parallelism between both languages is certainly wrong due to either the absence of text in one of the two languages, the interference between the text in the two languages without separators, or a big mismatch in the length of the two phrases or the number of punctuators.
Data cleaning deals with the problems of Unicode text-coding; like dropping non-alphabetic codes such as musical symbols, emoticons, control characters … etc. It also eliminates phrases with rogue words; like excessively long words, and eliminates diacritics from Arabic words. Data cleaning also unifies punctuators such that all the punctuators in different languages (Arabic, English … etc.) are all converted to English punctuators. It also inserts spaces between punctuators attached to words without spacing between them, and it also replaces each series of repeated consecutive punctuators by a single punctuator of the same type. Each series of over-repeated consecutive alphabetical characters is also fixed; e.g. "مبروووووووك"→"مبروك" .
Spell-checking is necessary to deal with raw text parallel corpora with modest text quality – like *OpenSubtitles* that we worked on in this paper – so that stark spelling errors and typos are handled. We built a simplistic spell-checker that handles such errors in English and Arabic; e.g. non-spaced consecutive words, spaces inserted in the middle of words, and confusing "أ" with "ا" and "ة" with "ه" and "ي" with "ى" … etc. in Arabic.

C. The Algorithm

After preprocessing the raw parallel text corpus, our algorithm *Bidirectional Semantic Augmentation* is ready to work on the pre-processed parallel text corpus as follows:

Step 1: We determine the two languages L1 and L2 that represent the two directions in the paraphrases dataset and covered by the parallel text corpus. In the work presented in this paper L1 is English and L2 is Arabic, without any loss of generality of the algorithm.

Step 2: We extract a bidirectional thesaurus from the pre-processed parallel text corpus. This bidirectional thesaurus is composed from two halves; a phrase thesaurus from L1 to L2 (in step 3) and a phrase thesaurus from L2 to L1 (in step 4).

Step 3: We extract the phrase thesaurus from L1 to L2. Each entry in this thesaurus is composed of a key phrase (from the corpus) in L1 and its corresponding phrases (from the corpus) in L2 – see figure 4. This extraction takes place through the following three sub-steps:

> Step 3.1: Unique phrases in L1 that appeared in the corpus are recognized, and each of these unique phrases becomes a key phrase in one of the phrase thesaurus entries.
>
> Step 3.2: For each key phrase in L1 we register its corresponding unique phrases (from the corpus) in L2. Here are two (simplified illustrative) examples from our work:
>
> {English key phrase: **Good morning**, Arabic corresponding phrases: {صباح الخير، عمت صباحا، صباحو
>
> {English key phrase: **Happy morning**, Arabic corresponding phrases: {صباحك سعيد، أسعد الله صباحك

Step 3.3: Adding the number of occurrences of the key phrase and the number of occurrences of each of its corresponding phrase completes the entry in the phrase thesaurus from L1 to L2.

Step 4: We extract the phrase thesaurus from L2 to L1. Each entry in this thesaurus is composed of a key phrase (from the corpus) in L2 and its corresponding phrases (from the corpus) in L1 – see figure 5. This extraction takes place through the following three sub-steps:

Step 4.1: Unique phrases in L2 that appeared in the corpus are recognized, and each of these unique phrases becomes a key phrase in one of the phrase thesaurus entries.

Step 4.2: For each key phrase in L2 we register its corresponding unique phrases (from the corpus) in L1. Here are two (simplified illustrative) examples from our work:

{Arabic key phrase: **صباح الخير**, English corresponding phrases: **Good morning, Morning, Good day**}

{Arabic key phrase: **صباحو**, English corresponding phrases: **Good morning, Happy morning**}

{Arabic key phrase: **عمت صباحًا**, English corresponding phrases: **Sunrise, Rise and shine**}

Step 4.3: Adding the number of occurrences of the key phrase and the number of occurrences of each of its corresponding phrase completes the entry in the phrase thesaurus from L2 to L1.

Step 5: We select the first key phrase - of any entry from the entries of the phrase thesaurus from L1 to L2 - that is marked as "*Not used as a seed*" and mark it as "*Used as a seed*". Then we build two lists; one is T1 for synonymous phrases in language L1 that is initialized with a key phrase marked as "*Not used as a seed*" from the phrase thesaurus from L1 to L2 and mark it and its entry in the thesaurus with a flag "*Not invoked yet*", and the other list is T2 for the synonymous phrases in language L2 that is initialized empty.

If there is no key phrases (of some entry from the entries of the phrase thesaurus from L1 to L2) marked as "*Not used as a seed*" is left, [3] jump to step 11.

Step 6: We select the first phrase on the list T1 whose flag is "*Not invoked yet*" and change it to "*Invoked*", and for this *key* phrase we invoke (from its entry in the thesaurus from L1 to L2) all its *target* phrases in language L2. These target phrases are added on the list T2, and we mark the added phrases that were not on the list T2 with the flag "*Not invoked yet*". If there is not any phrases marked with the flag "*Not invoked yet*" on the list T1 or T2, jump to step 9.

Step 7: We select the first phrase on the list T2 whose flag is "*Not invoked yet*" and change it to "*Invoked*", and for this *key* phrase we invoke (from its entry in the thesaurus from L2 to L1) all its *target* phrases in language L1. These target phrases are added on the list T1, and we mark the added phrases that were not on the list T1 with the flag "*Not invoked yet*". If there is not any phrases marked with the flag "*Not invoked yet*" on the list T1 or T2, jump to step 9.

Step 8: Go back to step 6.

Step 9: The elements of the list T1 together make a paraphrase in language L1, and the elements of the list T2 together make its corresponding paraphrase in language L2.

If we apply the previous steps from 5 to 9 on the simplified illustrative examples mentioned in sub-steps 3.2 and 4.2 starting with the key phrase "**صباح الخير**", we get the following English part of the paraphrase:

{**Good morning**, **Morning**, **Good day**, **Happy morning**, **Sunrise**, **Rise and shine**}

And the following Arabic part of the paraphrase:

{صباح الخير، عمت صباحًا، صباحو، صباحك سعيد، أسعد الله صباحك}

[3] All the key phrases in each entry of the phrase thesaurus entries with its two halves are marked as "*Not used as a seed*" upon the start of running the algorithm.

It should be noted that the numbers of occurrences of synonymous phrases are registered during the execution of steps from 5 to 9, and the phrases in each paraphrase part of language L1 are sorted in a descending order, and the most frequent phrase among them is called the *head phrase* in the paraphrase part of language L1. The same happens with the phrases in the paraphrase part of language L2 regarding the registration of the numbers of their occurrences and ordering these phrases according to their numbers of occurrences, and the most frequent phrase among them is called the *head phrase* in the paraphrase part of language L2.

Step 10: Go back to step 5.

Step 11: Exit.

4 STATISTICAL QUALITY EVALUATION

In order to judge how good our new *Bidirectional Semantic Augmentation* algorithm is, it is essential to evaluate the semantic quality of its outputs. The semantic quality of a bidirectional (English-Arabic) paraphrase is intuitively 100% perfect if "every one of the English phrases in the paraphrase has at least one interpretation such that all these English phrases are synonymous or *almost* synonymous" AND "every one of the Arabic phrases in the paraphrase has at least one interpretation such that all these Arabic phrases are synonymous or *almost* synonymous". Meanwhile, each phrase in the paraphrase that is not synonymous with the majority of the other phrases reduces the semantic quality of the paraphrase.
To turn the semantic quality from a *qualitative concept* to a *quantitative metric*, we followed the simple procedure upon inspecting a bidirectional/bilingual paraphrase:

```
{       Let L1 be short hand for: One of the languages pair.
        Let L2 be short hand for: The other one of the languages pair.

        Let SQ be shorthand for Semantic Quality of the bidirectional paraphrase.

        Let SQ1 be the SQ of the L1 part of the paraphrase.
        Let SQ2 be the SQ of the L2 part of the paraphrase.

        Let N1 be the count of all the phrases of L1 in the bidirectional paraphrase.
        Let NS1 be the count of those phrase of L1 that can be regarded synonymous with the majority of the N1 phrases.
        SQ1 = NS1 / N1

        Let N2 be the count of all the phrases of L2 in the bidirectional paraphrase.
        Let NS2 be the count of those phrase of L2 that can be regarded synonymous with the majority of the N2 phrases.
        SQ2 = NS2 / N2

        SQ = (SQ1 + SQ2) /2
}
```

Upon inspecting a phrase to judge whether it is synonymous (or almost synonymous) to the majority of the other phrases in its same language in the same paraphrase, we do not care much about its spelling, morphological or grammatical errors as long as it is comprehensible by a native speaker of the language. This sounds reasonable as we are performing a semantic judgment, not a morphological nor a syntactic assessment. Moreover, the raw corpus – like the one we used to produce the work presented in this paper – could be full of informal language.
To show how to apply the procedure presented above, we compute the semantic quality of the paraphrase illustrated in figure 6 below.

L1 here refers to English, and L2 here refers to Arabic

SQ1 is the SQ of the English part of the paraphrase, and
SQ2 is the SQ of the Arabic part of the paraphrase.

N1 = 9, NS1 = 9 (because all the English phrases are synonymous)
Therefore, SQ1 = NS1 / N1 = 9/9 = 1

N2 = 19, NS2 = 18 (one Arabic phrase - shaded by grey - is not synonymous to the other 18 Arabic phrases)
Therefore, SQ1 = NS2 / N2 = 18/19 = 0.947

The semantic quality SQ of the whole paraphrase = (SQ1 + SQ2) /2 = (1 + 0.947) /2 = 0.974 = 97.4%

The overall semantic quality index of the whole output of paraphrases dataset could simply be computed as the arithmetic average of the semantic quality indices of each paraphrase in the output dataset.

English phrases of the paraphrase		Arabic phrases of the paraphrase	
Number of synonymous English phrases = 9 **Sum of the frequencies of English phrases = 28**		**Number of synonymous Arabic phrases = 19** **Sum of the frequencies of Arabic phrases = 55**	
		6	أحسنت صنعا
		5	لقد أبليت حسنا
		4	فعلت خيرا
		4	لقد أبليت جيدا
		4	أحسنت
You did good	5	3	أبليت حسنا
You did fine	4	3	لقد ابليت حسنا
You did well	4	3	قمت بعمل جيد
You have done well	3	3	لقد قمت بعمل جيد
You done good	3	2	أبليت بلاء حسنا
You did great	3	2	ابليت حسنا
You did all right	2	2	لقد قمتي بعمل جيد
You did very well	2	2	هذا جيد
you did good	2	2	أنت لم جيدة
		2	أنت فعلت جيدا
		2	أحسنت عملا
		2	قمت بعمل رائع
		2	لقد قمتى بعمل جيد
		2	لقد فعلت أمر جيد

Figure 6: One of the English-Arabic paraphrases produced by our work used as a subject to show how to compute the semantic quality of a bilingual/bidirectional paraphrase

However, it would consume a prohibitively long time and a huge amount of human resources to manually inspect and compute the semantic quality of each one of the output 181,467 paraphrases.
Fortunately, the overall semantic quality index of the whole output of paraphrases dataset can be estimated – to an arbitrary degree of approximation - through the inspection of a random sample of the whole dataset. If we tolerate a confidence level of estimation less than 100% and a non-zero error margin in the estimated quality index, the size of the random sample to be inspected tends to be a small fraction of the size of the whole dataset when the size of the whole dataset is large and the average semantic quality index is close to 100%. [13]
For a confidence level of estimation of 95% along with an error margin of estimation of 2.21%, we were required to assign specialized linguists to manually inspect only 235 randomly selected paraphrases whose arithmetic average of measured semantic quality indices was found to be 96.92% [4]. Table 1 below summarizes this statistical semantic evaluation. [5]

TABLE 1
Summary of the statistical semantic evaluation of the paraphrases dataset produced by our Bidirectional Semantic Augmentation algorithm when run on the English-Arabic parallel corpus from the 2018 version of *OpenSubtitles* (OPUS)

Number of English-Arabic paraphrases in the whole output dataset (community size)	181,467
Estimation confidence level (confidence level)	95%
Error margin in the estimated average semantic quality index (error margin)	2.21%
Number of the randomly selected and manually inspected sample paraphrases (sample size)	235
Estimated overall Semantic Quality Index $\in$ [96.92% – 2.21%, 96.92% + 2.21%]	96.92%

5 Conclusion

In this paper we dealt with the problem of automatic bidirectional paraphrasing via pivoting on a large parallel text corpus, and mentioned few examples of its applications in the field of natural language processing. We detailed the structure of the paraphrases dataset we deliver, and we reviewed the previous works on "automatic paraphrasing starting from parallel text corpora" over the last two decades.
The paper then proceeded to present our novel *Bidirectional Semantic Augmentation* algorithm to handle that problem. We ran this novel algorithm on the English-Arabic text corpora from the 2018 version of the *OpenSubtitles* (OPUS) parallel text corpora, which represents a challenging vibrant everyday conversational language that both covers a wide range of genres and the contemporary era extending from 1930's to 2010's.
While this new algorithm is much simpler and more computationally efficient than the other existing methods in this regard, our random-samples based statistical evaluation of the semantic quality of its outputs indicates its interestingly superior performance.
The deliverables of our work are available for research purposes at:
https://github.com/FahadGhamdi/Bidirectional-Paraphrases-Dataset-from-an-English-Arabic-Subtitling-Parallel-Corpus

[4] So, the estimated arithmetic average of the semantic quality $\in$ [96.92% – 2.21%, 96.92% + 2.21%]

[5] These numbers were obtained using one of the most renowned and reliable online and interactive Sample Size Calculation tools: *https://www.calculator.net/sample-size-calculator.html*.

ACKNOWLEDGEMENT

This research is funded by the Translation Studies Grants Program of the Arabic Observatory of Translation (AOT) in KSA *https://aotalecso.org* to complete this study in the field of translation for the year 2023. Our research team thanks the Arab Translation Observatory for their generous support of this work.

REFERENCES

[1] Bannard, C., and Callison-Burch, C. (2005) "Paraphrasing with Bilingual Parallel Corpora." in the Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 25-30 June 2005, (ACL 2005) - University of Michigan – USA.

[2] Callison-Burch, C. (2008) "Syntactic Constraints on Paraphrases Extracted from Parallel Corpora." in the Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 25-27 October 2008, (EMNLP 2008), Honolulu – Hawaii - USA, A Meeting of SIGDAT, a Special Interest Group of the ACL.

[3] Creutz, M. (2008) "Open Subtitles Paraphrase Corpus for Six Languages." in the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018, Miyazaki - Japan.

[4] Denkowski, M. and Lavie, A. (2010) "METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages." in the Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, July 2010, Uppsala, Sweden, Association for Computational Linguistics pp. 339-432, https://aclanthology.org/W10-1751

[5] Dolan, B. and Brockett, C. (2005). "Automatically Constructing a Corpus of Sentential Paraphrases." in the Proceedings of the Proceedings of the Third International Workshop on Paraphrasing (IWP2005), https://aclanthology.org/I05-5002

[6] Dolan, B., Quirk, C., and Brockett, C. (2004) "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources." in the Proceedings of Proceedings of the 20th international conference on Computational Linguistics (COLING '04), Geneva –Switzerland, August 23 - 27, 2004, Association for Computational Linguistics.

[7] Ganitkevitch, J. and Callison-Burch, C. (2014) "The Multilingual Paraphrase Database" in the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) pp. 4276–4283, May 2014, Reykjavik - Iceland.

[8] Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013) "PPDB: The Paraphrase Database" in the Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013) pp. 758–764, June 2013, Atlanta – Georgia – USA.

[9] Lison, P., and Tiedemann, J. (2016) "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles." in the Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož - Slovenia.

[10] Manning, C., and Schütze, H. (1999) "Foundations of Statistical Natural Language Processing.", MIT press - USA.

[11] Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2015) "PPDB 2.0: Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification." in the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), July 2015, Beijing - China, pp. 425–430, https://aclanthology.org/P15-2070.

[12] Quirk, C., Brockett, C., and B. Dolan, W. (2004). "Monolingual Machine Translation for Paraphrase Generation." in the Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona – Spain.

[13] Ryan, T. (2013) "Sample Size Determination and Power", Wiley Publishing Company (John Wiley & Sons Inc.), New Jersey - USA.

[14] Tschirsich, M. and Hintz, G. (2013) "Leveraging Crowdsourcing for Paraphrase Recognition." in the Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Aug. 2013, Sofia – Bulgaria, Association for Computational Linguistics, https://aclanthology.org/W13-2, pp. 205 – 213.

[15] Tiedemann, J. (2016) "Finding Alternative Translations in a Large Corpus of Movie Subtitles." in the Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož - Slovenia.

[16] Zhou, J, and Bhat, S. (2021) "Paraphrase Generation: A Survey of the State of the Art." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5075-5086, Online and Punta Cana – Dominican Republic.

ARABIC ABSTRACT

استخراج وتوسيع قائمة مزدوَجة اللغة بمجموعاتِ العباراتِ المترادِفةِ من مدوَّنةٍ نصيَّةٍ إنجليزيةٍ–عربيةٍ متوازيةٍ من ترجمات الأعمال الدرامية

فهد الغامدي [١] عبد العاطي هواري[٢] محمد عطية [٣]

[١] أستاذ مساعد بقسم علوم الحاسب – جامعة الباحة – الباحة – المملكة العربية السعودية، fghamdi@bu.edu.sa

[٢] المؤسس والمدير التنفيذي لشركة DataLex4AI – سانتا كلارا – كاليفورنيا – الولايات المتحدة الأمريكية abdelati@datalex4AI.com

[٣] استشاري أبحاث في "الشركة الهندسية لتطوير النظم الرقمية RDI" – الجيزة – مصر Mohamed.Attia.NLP@gmail.com

ملخص

يعد تمييزُ مجموعات العبارات المترادفة واستخلاصُها من أكثر المهام الرئيسية التي تمثل تحدّيًا في مهام التحليل الدلالي العميق لِلُّغات الطبيعية بما له من أهمية قصوى في معالجة هذه اللغات ومن ثَمَّ في الذكاء الاصطناعي التوليدي. وفي هذه الورقة نعرض خوارزمية جديدة تُشَغَّل على مدونة ضخمة من النصوص المتوازية لأي زوجين من اللغات فتستخلص تلقائيًّا مجموعات العبارات المترادفة (أو المتقاربة جدًّا في المعنى) في هذين الزوجين من اللغات. وكما تعوّل بعض الخوارزميات السابق إنتاجها في هذا الصدد على الترجمة ثنائية الاتجاه التي توفرها المدونات النصية المتوازية الضخمة من أجل استخلاص أزواج العبارات المترادفة، فإن خوارزميتنا الجديدة تعول على هذه الميزة ذاتها، ولكنها خوارزمية أبسط وأكفأ من سابقاتها. وعلاوة على ذلك، فإن خوارزميتنا هذه تنفرد بقدرتها على استخلاص وبناء مجموعة العبارات المترادفة كاملةً نتيجةَ تشغيلِها دفعةً واحدةً على مدونة نصية متوازية بلا حاجة إلى أية معالجات لاحقة. وقد قمنا بتنفيذ خوارزميتنا وتشغيلها على المدونة النصية الإنجليزية–العربية ضمن إصدار عام 2018م لبنك المدونات النصية المتوازية الشهير لترجمات الأعمال الدرامية (OPUS) *OpenSubtitles*، ثم قمنا بقياس متوسط الجودة الدلالية لمجموعات العبارت المترادفة المستخلصة تلقائيًّا قياسًا إحصائيًّا عن طريق فحص عينات مختارة عشوائيًّا، وتبين من ذلك أن متوسطَ الجودةِ الدلالية المتحققَ مرتفعٌ على نحو ملفت للنظر.

كلمات مفتاحية

تحليل دلالي، ترادُف، عبارة، مجموعة عبارات مترادفة، استخلاص مجموعات العبارات المترادفة، مدونة نصية متوازية.