Evaluating Neural Machine Translation Using Error Analysis In English -Arabic Texts

By

Fahad Saad Alsahli

Assistant Professor, Department of English

College of Science and Humanities

Prince Sattam Bin Abdulaziz University

Kingdom of Saudi Arabia

تقييم الترجمة الآلية العصبية بإستخدام تحليل الأخطاء في ترجمة النصوص من اللغة الإنجليزية إلى اللغة العربية

فهد بن سعد السهلي – أستاذ مساعد – قسم اللغة الانجليزية
كلية العلوم والدراسات الانسانية–جامعة الأمير سطام بن عبد العزيز
المملكة العربية السعودية

ملخص:

هدفت هذه الدراسة إلى تقييم مخرجات الترجمة الآلية العصبية لترجمة النصوص من اللغة الإنجليزية إلى اللغة العربية باستخدام منهجية تحليل الأخطاء. أستخدم موقع ترجمة جوجل في هذ الدراسة حيث يمثل أحد أبرز المواقع التي تعتمد على الترجمة الآلية العصبية. معظم الدراسات التي أجريت على الترجمة الآلية أجريت على برامج ومواقع تستخدم نظم الترجمة الآلية الإحصائية أو القائمة على القواعد بدلاً من الترجمة الآلية العصبية. تم اختيار النصوص بناءً على معايير جمعية المترجمين الأمريكية المستخدمة في اختباراتهم. تم اختيار ثلاثة نصوص لتمثيل ثلاثة أنواع مختلفة من النصوص وهي: النصوص العامة والمالية والعلمية. ثم استخدمت منهجية تحليل الأخطاء لتحليل نتائج الترجمة ومقارنتها مع بعضها البعض ومع ورد في الدراسات السابقة. يتضح من هذه الدراسة أن هناك ١٠٥ أخطأء في النصوص الثلاثة بمعدل ١,٩ خطأ لكل جملة. ٢٧ خطاء من الأخطاء كانت أخطاء نحوية ، في حين أن ١٤ من إجمالي الأخطاء هي أخطاء متعلقة بقواعد اللغة ، و ٦٤ من الأخطاء هي أخطاء دلالية. على الرغم من وجود تحسن واضح في ترجمة جوجل ، بعد تحويله لنظام ترجمة آلي عصبي ، خاصة في جزء القواعد ، إلا أنه يجب عمل المزيد لتحسينه بشكل عام وفي الجزء الدلالي بشكل خاص.

الكلمات المفتاحية : الترجمة الآلية العصبية ، تحليل الأخطاء، الترجمة، تقييم الترجمة، تقييم الترجمة الآلية العصبية.

Evaluating Neural Machine Translation Using Error Analysis In English -Arabic Texts

Abstract

The aim of this study was to evaluate the output of Neural Machine Translation of translating texts from English into Arabic using error analysis. Google Translate was taken as an example as the leading neural machine translations. Most of the studies done on machine translation were on rule-based and statistical machine translation rather than neural machine translation. Texts were selected based on the American Translator Association criteria which is used in their examinations. Three texts were selected to represent three types of texts: general, financial, and scientific. Error analysis then was used to analyze the results of the translation and compare them with each other and with that in the literature. 105 errors were discovered in the three texts with an average of 1.9 error per sentence. 27 of the errors were syntactic errors, while 14 of the total errors are grammatical errors, and 64 of the errors are semantic errors. Although there is a clear improvement in Google Translate ,especially in the grammar part , since it was shifted to a neural system, more has to be done to improve it in general and in the semantic part in particular.

Keywords Neural machine translation, Error analysis, Translation, Translation evaluation, Machine translation evaluation.

1. Introduction

It has been almost a century since the real thinking of machine translation has started, yet the status of the machine translation seems to be less than what most of the scientists were expecting at the beginning of the research on it. It is important to note that the development of machine translation seems to go hand in hand with the development of technology since it depends on it. Therefore, there is no wonder that the pace of development in machine translation has accelerated in the last few decades. One of the big improvements in technology is the widespread use of internet, which seems to change the whole front of machine translation. For the first time in its history machine translation has become available for professional and freelancer translators either free or for a reasonable price. This led many of them to adopt this technology which results in big improvement in it to satisfy the users. One of the most available and widely used machine translation is Google Translate since it is free and can be accessible from any part in the world as long as there is an internet connection. Google in its part has realized the importance of its translation system and worked continuously on improving it. At the beginning they start with a statistical machine translation which was the norm at that time. Their system, and because of its popularity, was subject to many scientific researches with the aim of improving it. With the development of deep learning, and neural machine translation, Google Translate has decided to shift its system to a new system based on these new technologies. Google has claimed that the new system has gained improvement in one night and it now equals the improvement of the old system over its entire life.

2. Problem statement

Although there are many researches on Google Translate available in the literature, most of these researches are done on the old system of Google Translate; the statistical-based system. Google has made a big claim about the quality of the new system which attracts some researchers to analyze the new system for different language pairs. So far, it seems that there are very few studies tackling the new system of Google Translate in translating document from English into Arabic. More importantly most of the available studies in the literature have tackled Google Translate at the word, phrase, or

sentence levels, and only very fewer studies have tackled it at the text level.

3. Research Objectives

The specific objectives of this study are as follows:

1. To identify the errors made by Google Translate as a neural machine translation when translating texts from English into Arabic.
2. To classify these errors according to a well-established frame of analysis based on error analysis.
3. To compare the errors made across their types and across the text types as well.

4. Literature Review

4.1. Machine Translation

Machine translation can be defined as " the process by which a computer software is used to translate a text from one natural language to another (Al Humaidan, 2001; Karami, 2014). Some researchers argue that the concept of machine translation is dated back to the 17th century when J.J.Becher wrote a dictionary in which 10000 Latin words were matched with digit numbers so they can be easily replaced by their correspondence words in different languages (Al Humaidan, 2001) . However, the real beginning of thinking about machine translation can be traced back to the mid-1930s when both Artsrouni and Troyanskii applied for patents for "translating machines" in France and the Soviet Union respectively(Mohammed, Samad, & Mahdi, 2018). Artsrouni concept about machine translation was not so developed as Troyanskii's who put a complete proposal containing outline for coding interlingual grammatical rules and guidelines on how the analysis might work, in addition to the use of electronic bilingual dictionaries (Al Humaidan, 2001). However, these were mere proposals of the concepts itself and far from being applicable in real life. But when the computer was invented and the WWII had just finished, there was a real and an urgent need for machine translation for intelligence purposes. In 1949, Warren Weaver wrote a memorandum containing various proposals on machine translation based on code breaking during the WWII. Within few years the research on machine translation was on a full scale in

different institutions in USA which led to the first public demonstrations of machine translation.

Although the first generation of systems were mainly large bilingual dictionaries, they seemed to raise the bar of developing a fully automated machine translation that produced high quality translation similar to the one produced by human translators. This led to a surge in the funding provided for research on machine translation from various agencies in the USA. However, most of the systems created during this era were below the expectations, which led the government to set up a committee called the Automic Language Processing Advisory Committee (ALPAC) to study the current state of machine translation and the future prospect of it. In 1966 the committee issued a famous report concluding that machine translation was far more expensive and far less effective than human translators. This report has almost brought the research on machine translation to standstill not only in the USA but also in other countries as well for almost a decade. Yet some companies in the USA and other countries continued to work on developing machine translation especially for intelligence and civil defense purposes where the need is on getting a raw translation more than a polished one. Researchers in this era also began to tackle the issue from different perspectives such as focusing on what is called sublanguages – i.e. languages used in specific contexts such as weather forecast – rather than one size fits all. This led to some success on some fronts such as the program created for weather forecast in the University of Montreal in Canada which was a successful machine translation program.

However, the second wave of machine translation did not emerge until the 1980s where the microcomputers and text-processing software became widespread. The 1990s witnessed a turning point in the history of machine translation with experiments based on purely statistical methods and corpora of translation examples. It also saw the beginning of speech translation research. The research on statistical and example-based machine translation continued to grow in the early 2000s with the help of the widespread use of translation aids, sales of machine translation for personal use, and the availability of online translation services. The widespread use of the internet led some giant companies such as Google to enter the world of machine translation. In 2006 Google launched its first machine

translation program called Google Translate which was based on statistical machine translation. The results of the translation were not very impressive. However, in 2016 Google shifted its machine translation to a new system based on "Neural Machine Translation". This new system is based on what is called deep learning. The machine in this system learns where to look for patterns in the data and learn from them to make classifications and predictions (Ducar & Schocket, 2018).

With this shift the number of the languages covered by this service increased to reach almost 105 language pairs. The quality of the translation also improved dramatically to the extent that the improvement achieved by this new system over one night was equal to the total improvement of the old system over its entire life (Lewis-Kraus, 2016). The new system reduced the errors by almost 60% based on human evaluations (Ducar & Schocket, 2018; Wu et al., 2016) .

As for the machine translation that tackled Arabic language, the first one was a machine translation to translate between English and Arabic and it is dated back to the 1970s where a professor at Harvard university developed a program for this purpose. However, this attempt was not successful due the big differences between the two languages (Madkour, 2011). Following that, different attempts were made to develop machine translation programs to translate between English and Arabic, and some of them were very successful such as " Almutarjim Al-Arabi" which was developed by a London-based company called ATA (Al-Samawi, 2014). This program is still available today and different versions of it have been developed as well such as "Alwafi" and "Almisbar".

4.2. Evaluation of Machine Translation output

There are two main approaches for the evaluation of machine translation output that are frequently used by researchers. The first one is the automatic evaluation of machine translation by using metrics that compares the machine translation output with a human reference translation and provides a score for the evaluation. The reason behind adopting such an approach is that it is cheaper and faster than the human evaluation. The most widely used of such an approach is the BLUE method which is developed by Papineni et al (2002). Different methods are used in this approach such as:

assessing the adequacy of the translation, measuring the amount of post editing needed for the output to be of an acceptable quality, error analysis, and content analysis.

Although the apparent advantage of this approach is that it is cheaper and faster to administer, some researchers argue that it is not an accurate measure and thus it is not sufficient to define the quality of machine translation (Ghasemi & Hashemian, 2016). The scores of the automatic evaluation does not necessarily reflect the quality of the translation (Callison-Burch, Osborne, & Koehn, 2006). Moreover, some scholars argue that since automated quality evaluation of machine translation compares the output of it with reference human translation, and then it only measures the similarity of the texts at superficial level. Therefore it conflates the fluency of form with the accuracy of the content (Koponen, 2010). Also, some researchers argue that human evaluation is the ultimate one and cannot be ignored (Papineni, Roukos, Ward, Henderson, & Reeder, 2002). Therefore researchers who want to use some of the automated methods of machine translation evaluation are advised to exercise cautions while doing that (Culy & Riehemann, 2003).

4.3. Evaluation of English-Arabic, Arabic-English machine translation output

Chalabi (2000) tested a free machine translation system to translate from English into Arabic and claimed that it could achieve an accuracy level of 60%. Al-Kabi et al have done a study in which they compared Google Translate with Babylon machine translation system (Al-Kabi, Hailat, Al-Shawakfa, & M, 2013). They used several sentences from various sources and translated them from English into Arabic. They then evaluated the results of the translation using BLEU. The results showed that Google Translate is better than Babylon. The two systems were compared again in another study by Kadhim et al, but this time news headlines were used to test the translation of both systems from English into Arabic (Kadhim, Habeeb, Sapar, Hussin, & Abdullah, 2013) . Again, BLEU was used to evaluate the two systems. Both studies concluded that Google Translate performed a little better in accuracy while Babylon in style, both systems had similar score of clarity.

Another study was done by Adly and Ansary to evaluate the Universal Network Language UNL translation against another well-known translation systems such as Google Translate and Babylon (Adly & Al Ansary, 2009). They also used BLEU to evaluate different systems. Their results showed that Google Translate performed better than other systems. ElShiekh carried out a study to evaluate Google Translate in translation from English into Arabic and vice versa (ElShiekh, 2012). Using three different types of texts as a basis for his evaluation, ElShiekh advised against taking Google Translate as the final translation. A similar study was carried out by Al-Samawi to evaluate Google Translate in translating encyclopedia texts (Al-Samawi, 2014). The study examined 10 different types of texts, each with 10 sentences, subjected to error and content analysis to facilitate the evaluation of the system. The evaluation was based on human evaluation rather than an automated evaluation. The results showed that there was an average of 3.66 error per sentence.

5. Method

5.1. Research procedure

This research is a descriptive study analyzing the output of Google Translate from three different disciplines. The three disciplines were selected following the guidelines of the American Translator Association (ATA) which is one of the largest translators' associations in the world (Koby & Champe, 2013). The ATA runs one of the most comprehensive certification examinations for translators in 29 language pairs. These three texts represented General, Financial, and Scientific disciplines. These three texts were inserted separately into Google Translate and the results of the translations were obtained for content and error analysis. Descriptive analysis of the translation was done to facilitate further analysis and to determine which texts contained most of the errors, and which types of errors were prevalent in the three texts. This descriptive analysis included things such as number of words and sentences in both the source and the target texts. A classification of errors developed by Al-Samawi (2014) was used to analyze the results of the translation. Al-Samawi developed his classification of errors based on procedures recommended by the some of the well-known scholars in error analysis such as Pit Corder. The results of the analysis were then presented to three linguists to determine the credibility of the classification of errors. No major disagreement was

found among the three linguists about the classification of errors. The results of descriptive analysis were compared to what was available in the literature to give the reader a more comprehensive picture about the status of Google Translate as one of the leading neural machine translations.

5.2. Data set

The three texts were chosen in accordance with ATA guidelines to represent three different types of texts; namely general, financial, and scientific (Koby & Champe, 2013). The general text was an editorial text from the Washington Post newspaper and consisted of 403 words. The financial text was an annual report published by The Bank of America and consisted of 399 words. Finally, the scientific text was about diabetes and was published by The American Diabetes Association. It consisted of 418 words.

5.3. Instrument

After selecting the texts to be used for this research, they were entered into Google Translate, and the outputs were obtained for analysis. Different categories for error analysis were prepared based on previous studies, and on Al-Samawi study's in particular. Table no1 shows the categories used for error analysis in this study.

Table 1 *Error Analysis Framework* (Al-Samawi, 2014)

Error Category	No	Type of Error
Syntactic Errors	1	Starting with a nominal sentence in place of a verbal sentence.
	2	Violating the whole phrase structure (Putting adjective before noun, Putting modifiers before modified terms)
	3	Using wrong form of the word (plural, the five verbs, five nouns, nouns and verbs inflections)
	4	Violating subject-verb agreement (masculine and feminine; singular, dual, and plural; first, second, and third person)

Grammar Errors	5	Using a noun in place of a verb
	6	Using a verb in place of a noun
	7	Using wrong prepositions, articles, and particles
	8	Using definite article before genitives
	9	Omitting functional morphemes (i.e., prepositions, articles, conjunctions, pronouns, auxiliary verbs, deixis, etc.)
Semantic Errors	10	Using a wrong meaning of English homonyms
	11	Using words of ambiguous meaning
	12	Using terms that convey very different meaning
	13	Using unfamiliar words in place of collocations
	14	Using wrong reference and relative pronouns.
	15	Adding an unnecessary word, preposition, or article before a word
	16	Omitting necessary words or phrases
	17	Corrupting the meaning of the whole sentence

6. Data Collection and Analysis

After deciding on the three texts used for this study, they were put into Google Translate and the results of the translation were obtained for analysis. The researcher went through the texts sentence by sentence and words by words identifying the errors in translation and labeled them according to the classification table. Descriptive analysis is presented in the following table.

Table 2 Descriptive Statistics of Source Texts and Their Translations

Text	source text		Translation	
	No of sentences	No of words	No of sentences	No of words
General	14	403	15	373
Financial	21	399	22	383
Scientific	17	418	17	438

An analysis of table no 2 shows that while there was an increase in the number of sentences for general and financial texts when translated into Arabic, the number of words for the whole texts decreased. But when it comes to the scientific text, the number of sentences were the same for both the source and the translated text, but the number of words in fact increased when translated into Arabic.

Table no 3 shows texts types and the number of errors made in the translation.

Table 3 Text Types and Number of Errors

Text	Total number of sentences	Total number of words	Total number of errors	Percentage of errors to total number of words	Percentage of errors in each text to all errors in all texts.
General	15	373	29	7.7%	27.6%
Financial	22	383	62	16.1%	59.0%
Scientific	17	438	14	3.1%	13.3%
Total	54	1194	105	-	100

Total number of errors is 105 compared to total number of sentences 54, that is 1.9 error for each sentence. The financial text scored the largest number of errors, that is 62 errors in 22 sentences, which means there is an average of 2.8 errors in each sentence. This represent 59.0% of the total errors in all of the three texts. The general text scored 29 errors in 15 sentences with a 27.6% of all errors in the three texts and an average of 1.9 errors in each sentence of the text. The scientific texts scored the least of errors with 13 errors in 17 sentences, that is an average of 0.8 error in each sentence, and a total of 13.3 % of all errors in the three texts. The most common errors in all texts was error no 15 (Adding an unnecessary word, preposition, or article before a word) with 16 errors in 54 sentence and a 15.2% of all errors. The least common errors are error no 3 (Using wrong form of the word (plural, the five verbs, five nouns, nouns and verbs inflections) and error no 6 (Using a verb in place of a noun) which represents 0.9% of all errors. The most common types of errors are the semantic errors with a total of 62 errors and a percentage of 59.0% of all errors. In the second place comes the syntactic errors with 27 errors which represents 25.7% of all errors in all texts. The least common types of errors were the grammar errors with 14 errors which represented 13.3% of all errors.

7. Results and Discussion

The main objective of this study was to evaluate the translation of a neural translation system from English into Arabic. Google Translate was taken as an example of this kind of machine translation. Three texts were selected and used to allow for analysis across different kinds of genres. The output of the translation of these three texts were then analyzed through error analysis. A category of error analysis developed by Al-Samawi (2014) was used in this study. The most common type of errors found in this study was error type 15 (27.6%). This almost corresponds to what Al-Samawi found in his study where error type 15 came as the second most common type of error after error type 9 (13.9%). The least common error found in this study were errors type 3 and 6. Again this seems to replicate Al-Samawi's results where error type 6 was the least common one. The most common type of errors in this study was the semantic errors while in Al-Samawi's study the most common one was the grammatical errors. However, in this study the grammatical errors were the least common type of errors. The syntactic errors came in the second place while in Al-Samawi's study they were the least common errors. It is

important to note that although Al-Samawi's study was done on Google Translate as well, Google Translate at that time was based on statistical machine translation system.

From the tables mentioned in the analysis section, it seems that Google Translate handles the scientific texts very well with an average of 0.8 errors in each sentence. This means one can get a fair translation of scientific texts from Google Translate which might need a minimum editing or even no editing at all. This might be due to the fact that scientific language is used in a very specific and direct way. Since neural machine translation is based on deep learning and can learn from the context, it seems that it can identify the text genre. Therefore, it is able to do a very good job in translating scientific texts since its language is simple and direct. When it comes to general texts, it seems that Google Translate is not able to perform as well as it does with the scientific texts. This is understandable as the general language is more complicated and sometimes it carries hidden meaning such as figure of speech. With an average of 1.9 errors per sentence, the output of the translation is fairly readable and can give the general meaning or the gist of the text. To get a good translation there must be a fair amount of editing to the translation output. When it comes to the financial texts, Google Translate performs the least with an average of 2.8 errors per sentence and a percentage of 59.0% of all errors in the three texts. This represents more than half of the errors in the translation output of all of the three texts. This comes as a surprise as one would assume that Google Translate performs better when it comes to specific language genres as it did with the scientific text in this study. However, in this study the translation of the general texts produced by Google Translate was better than the translation produced of the financial texts.

To help the reader gains more understanding of the kinds of errors committed by Google Translate, examples of the three classifications of errors are presented here.

Syntactic errors

There are 27 syntactic errors in the three texts which represents 25.7% of the total errors. One of the errors in this category is error type 1 "Starting with a nominal sentence in the place of a verbal sentence". An example of this error appeared in the financial text where the sentence reads "Economic growth was supported by a

noticeable pickup in business investment". In this example Google Translate rendered this sentence as " إقتصادي كان النمو مدعوماً بالتقاط ملحوظ" starting with a nominal sentence. The correct translation should start with a verbal sentence such as "كان النمو الإقتصادي". The other error in this category is error type 2 "Violating the whole phrase structure (Putting adjective before noun, Putting modifiers before modified terms). An example of this error is in the translation of the general text. The original phrase in the sentence is "the Democratic National Committee". The machine rendered this as " الديمقراطي اللجنة الوطنية" putting the modifier before the modified term. The correct translation should read "اللجنة الوطنية الديمقراطية".

Grammar errors

In this class of errors, Google Translate made 14 errors in all the three texts, which represents 13.3% of all errors. One of the errors in this category was error type 9 "Omitting functional morphemes (i.e., prepositions, articles, conjunctions, pronouns, auxiliary verbs, deixis, etc.)". An example of this error appeared in the translation of the scientific text. The original sentence is "Type 1 diabetes and type 2 diabetes are heterogeneous diseases in which clinical presentation and disease progression may vary considerably". In this example, Google Translate translated it as " مرض السكري من النوع ١ ومرض السكري من النوع ٢ من الأمراض غير المتجنسة التي قد يختلف العرض السريري وتطور المرض بشكل كبير". The correct translation should add the preposition "فيها" before "بشكل كبير", and the whole sentence should read like " مرض السكري من النوع ١ و مرض السكري من النوع ٢ من الأمراض غير المتجانسة التي قد يختلف العرض السريري وتطور المرض فيها بشكل كبير". Another error in this category was error type 7 "Using wrong prepositions, articles, and particles". An example of this error is seen in this sentence which is from the general text, "He approached a former senior Republican official for advice". In this example Google Translate translated it as "اتصل مع مسؤول جمهوري كبير سابق للحصول على المشورة", while the correct preposition should be "اتصل بمسؤول". A third error in this class of errors was error type 6 "Using a verb in place of a noun". An example of this kind of errors appeared in the financial text. The original sentence was "Following a midyear decline, long-term Treasury yields recovered towards the end of 2017". In this example, the machine rendered the sentence as " بعد منتصف العام تراجعت عائدات سندات الخزانة طويلة الأجل" treating the word "decline" as a verb and

translating it as " تراجعت" while it should be a noun, and should be translated as " تراجع".

Semantic errors

This is the most prevalent class of errors committed by Google Translate. 64 out of a total of 105 errors were from this category of errors. This represents about 60.9% of all errors discovered in the translation of the three types of texts in this study. One of the types in this classification was error type 15 "Adding an unnecessary word, preposition, or article before a word" which is the most common error identified in this study. An example of this type of error is seen in the translation of the scientific text where the sentence reads "General treatment goals and guidelines, and tools to evaluate quality of care". Google Translate has rendered this as " وأهداف العلاج العامة والمبادئ التوجيهية والأدوات اللازمة لتقييم جودة الرعاية" adding the word "المبادئ" to the translation of "guidelines". The translation of this word should be "توجيهات". In the same sentence Google Translate added the word "اللازمة" to the translation of "tools". But the correct translation should be " الأدوات" and there is no need to add "اللازمة" to the translation.

Another type of errors in this class was error type 16 "Omitting necessary words or phrases". An example of this type of error can be seen in the translation of the general text in this sentence "We feel confident that it was in that same spirit that Mr. Taylor agreed to testify about Mr. Trump's extortion of the Ukrainian government". This time Google Translate rendered this sentence as " إننا نشعر بالثقة من أن السيد تيلور وافق بنفس الروح على ابتزاز السيد ترامب للحكومة الأوكرانية" omitting the translation of the word "testify about" which resulted in the whole sentence giving the opposite meaning. So it became as if Mr.Taylor had approved Mr.Trump's actions in extorting the Ukrainian government. A third error in this category was error type 10 "Using a wrong meaning of English homonyms". For example, in the financial text Google Translate translated this sentence ""Economic growth was supported by a noticeable pickup in business investment in high-tech equipment, a recovery in oil exploration and solid consumer demand growth" as " اقتصادي كان النمو مدعوماً بالتقاط ملحوظ في الأعمال الاستثمار في معدات التكنولوجيا الفائقة، وانتعاش في التنقيب عن النفط ونمو قوي في الطلب على السلع الاستهلاكية". Google Translate translated "Business investment " into "الأعمال الإستثمار". In this case putting the modifier before the modified term (error type 2) and using a wrong

meaning of the English homonyms "business" (error type 10). The correct translation should read " الإستثمار التجاري".

8. Conclusion

The present study evaluated neural machine translation by identifying errors made by one of the leading neural machine translation system. A framework of error analysis developed by Al-Samawi was used in this study. This allowed the researchers to make some comparison between statistical machine translation and neural machine translation. Although there seemed to be an improvement in Google Translate as it moved to the neural machine translation, there are still some deficiencies in its translation from English into Arabic. The average of error per sentence seemed to improve from 3.66 in Al-Samawi's study to 1.9 error per sentence in this study. This seems to give more credibility to Google claim that its new translation system has scored an improvement of 60% over the old system (Ducar & Schocket, 2018; Wu et al., 2016).

It is important to note that this research was done on type of machine translation; Google Translate as an example of a neural machine translation. Although research is done on three types of texts, attention must be paid to the fact that texts from even the same type differs in their characteristics and level of complexity. Therefore, the results of the analysis cannot be generalized to cover all texts under these three types. It is hoped that this research contributes to the body of research on machine translation in general and neural machine translation in particular. Also, it is hoped that this research will contribute to the research on English Arabic machine translation as there is an urgent need for research on this particular pair of language.

References

Adly, N., & Al Ansary, S. (2009). Evaluation of Arabic machine translation system based on the Universal Networking Language. *International Conference on Applications of Natural Language to Information Systems*, *5723 LNCS*, 243–257. https://doi.org/10.1007/978-3-642-12550-8_20

Al-Kabi, M. N., Hailat, T. M., Al-Shawakfa, E. M., & M, I. (2013). Evaluating English to Arabic Machine Translation Using BLEU. *International Journal of Advanced Computer Science and Applications*, *4*(1), 66–73.

Al-Samawi, A. M. (2014). Language Errors in Machine Translation of Encyclopedic Texts from English into Arabic. *Arab World English Journal*, (3), 182–211.

Al Humaidan, A. (2001). *An Introduction to Machine Translation*. Riyadh: AL-Obiekan.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy: Association for Computational Linguistics.

Culy, C., & Riehemann, S. Z. (2003). The Limits of N-Gram Translation Evaluation Metrics. *System*, 71–78. Retrieved from https://docs.google.com/viewer?url=http://www.mt-archive.info/MTS-2003-Culy.pdf

Ducar, C., & Schocket, D. H. (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google translate. *Foreign Language Annals*, *51*(4), 779–795. https://doi.org/10.1111/flan.12366

ElShiekh, A. A. A. (2012). Google Translate Service: Transfer of Meaning, Distortion or Simply a New Creation? An Investigation into the Translation Process & Problems at Google. *English Language and Literature Studies*, *2*(1), 56–68. https://doi.org/10.5539/ells.v2n1p56.

Ghasemi, H., & Hashemian, M. (2016). A Comparative Study of Google Translate Translations: An Error Analysis of English-to-Persian and Persian-to-English Translations. *English Language Teaching*, *9*(3), 13. https://doi.org/10.5539/elt.v9n3p13

Kadhim, K. A., Habeeb, L. S., Sapar, A. A., Hussin, Z., & Abdullah, M. M. R. T. L. (2013). An evaluation of online machine translation of arabic into english news headlines: Implications on students' learning purposes. *Turkish Online Journal of Educational Technology*, *12*(2), 39–50.

Karami, O. (2014). The Brief View on Google Translate Machine. In *Seminar in Artificial Intelligence on Natural Language*. German.

Koby, G. S., & Champe, G. G. (2013). Welcome to the real world: Professional-level translator certification. *The International Journal for Translation and Interpreting Research*. Retrieved from http://search.informit.com.au/documentSummary;dn=282701995944286;res=IELHSS

Koponen, M. (2010). Assessing Machine Translation Quality with Error Analysis. *Electronic Proceedings of the VIII KäTu Symposium on Translation and Interpreting Studies*, *4*, 1–12.

Lewis-Kraus, G. (2016, December). The Great A.I. Awakening. *The New York Times Magazine*. Retrieved from https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html?mcubz=1

Madkour, O. (2011). Machine Translation - its concept and methods. *Journal of Dar AlUlom College*, (26), 893–937.

Mohammed, O., Samad, S., & Mahdi, H. (2018). A Review of Literature of Computer-Assisted Translation. *Language In India*, *18*(49042), 340–360.

Papineni, K., Roukos, S., Ward, T., Henderson, J., & Reeder, F. (2002). Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. *Proceedings of the Second International Conference on Human Language Technology Research*, 132–137. Retrieved from

http://portal.acm.org/citation.cfm?id=1289272

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 1–23. Retrieved from http://arxiv.org/abs/1609.08144