

AN EVALUATION OF ONLINE MACHINE TRANSLATION OF ARABIC INTO ENGLISH NEWS HEADLINES: IMPLICATIONS ON STUDENTS' LEARNING PURPOSES

Kais A. Kadhim, Ph.D

English Department, Faculty of Language and Linguistics, University of Malaya.
kaisamir@um.edu.my

Luwaytha S Habeeb

English Department, Faculty of Language and Linguistics, University of Malaya.

Ahmad Arifin Sapar, Ph.D

Arabic Department, Faculty of Language and Linguistics, University of Malaya
arifin@um.edu.my

Zaharah Hussin, PhD

Department of Educational Foundation and Humanities, Faculty of Education, University of Malaya
zaharah@um.edu.my

Muhammad Muhammad Ridhuan Tony Lim Abdullah

Management & Humanities Department, Universiti Teknologi Petronas, Malaysia
ridhuan_tony@petronas.com.my

ABSTRACT

Nowadays, online Machine Translation (MT) is used widely with translation software, such as Google and Babylon, being easily available and downloadable. This study aims to test the translation quality of these two machine systems in translating Arabic news headlines into English. 40 Arabic news headlines were selected from three online sources, namely *Aljazeera*, *daralhayat*, and *Aawsat*, where their English manually-translated versions were available. The selected data was evaluated by conducting criteria of Hutchins and Somers (1992) to find the assessment of each system outputs. Besides that, the selected data was also examined to find the types of translation techniques that are available in both machine outputs. A questionnaire was assigned to experienced professionals to evaluate the outputs to examine and determine which system was better to use in translating the collected data. The evaluation was based on criteria proposed by Hutchins and Somers. The findings indicated that both Google and Babylon had 80% of clarity, and Google scored a higher value of accuracy, i.e. 77.5%, compared to 75% of accuracy for Babylon. However, Babylon scored a higher value for style, i.e. 72.5%, compared to a score of 70% by Google. Nevertheless, the results revealed that online MT is undergoing improvement, and it has the potential to be one of the elements of globalization. As implication, the students could use online MT for learning purposes easily and quickly.

Keywords: MT, News Headlines, Google and Babylon translation, quality, and online MT Evaluation

INTRODUCTION

Researchers in the field of natural languages have undertaken a serious effort to support manual translations by inventing machine translations. Hutchins (1986, p: 15) defines Machine Translation (MT) as “the application of computers in the translation of texts, from one natural language into another”. Also known as automatic translation, MT has also been considered in the last decade as a computational linguistic phenomenon.

Apparently, MT is considered as a worthwhile subject for researchers, commercial developers and users (Hovy et al. 2002). As for researchers, they need to apply their theories to find out the differences that might be made by the machines. By doing so, it will be easier for developers to detect the most problematic issues and make the implementations on the system design. Evidently, the motive of commercial developers is to attract customers to buy their products. In turn, the users, who are interested in benefitting from MT, will decide which product meets their requirements. Examples of past researches and studies include the employment of various approaches to MT, such as studies by Marcu (2001), Richardson et al (2001), Tahir et al. (2010), and Groves (2006). Earlier researches focused on the direct approach such as the word-by-word analysis of the source language. Later on, researchers moved to the rule-based and statistical approaches. Salem (2009) is an example of this research trend. Meanwhile, there were researchers who were interested in the evaluation of MT quality since the users' demand increased for the use of machines with high levels of translation quality according to the rapid growth of technology and information. Different methods have been employed in measuring the quality of MT outputs according to different criteria outputs, such as *Fluency* and *Fidelity* (Eduard Hovy et al. 2002, p. 45). Some researchers analysed MT outputs for different purposes focusing on specific features; for instance,

agreement of number, and relative clauses (Flanagan, 1994). Others used the judgment of evaluators to rate whole sentences in terms of the N-point scale (White et al., 1992, 1994; Doyon et al., 1998), while others made use of the “bigram or trigram language model of ideal translation” to automatically measure the confusion which resulted from complexities in the target text (Papineni et al. 2001).

Schiaffina and Zearof (2005) at the 46th ATA Conference explained how translation quality could be measured, and they introduced “types of errors” in an output, whether these errors lie in meaning, form, or in compliance. However, their point of view of “good” translated output is to have zero errors, and their definition of quality does not differ from the main idea of all previous definitions. They defined it as “consistently meeting the needs and expectations of the customer or user”. Furthermore, they identified two categories of methods for evaluating the quality of translation, such as: “argumentative-centred systems”, and “quantitative-centred systems”. The former focuses on the functional relations between parts and whole, whereas the latter focuses on counting errors. An updated model of “argumentative-centred systems” was proposed by William (2009, p. 3-23), for assessing quality of translation.

The fact is that there are several methods of evaluating machine translation, which have been utilized to assess the outputs of translation. Round-trip is an example of these methods. Although this method seems to be good for evaluation, it has been described as “a poor predictor of quality” (Hutchins & Somers, 2005). The second example of evaluation methods is the human evaluation. The idea of this method is to train human for the purpose of translation assessment. The assessment based on comparing the various levels of human translation with machine translation output by making use of the judgements by human subjects. A good example is a study, which has been reported by Automatic Language Processing Advisory Committee (ALPAC), which tackles the comparison of different levels of machine and human translation from Russian into English based on two criteria, “intelligibility” and “fidelity”(ALPAC, 1966). Interestingly, Abraham and Salim (2005) proposed a model, based on shifts of structure and semantic. This model could be applied to evaluate MT outputs. Those shifts are either shifts of structure or shifts of semantic of the target language, (Cyrus, 2009, p. 103). Furthermore, in the same context of evaluation, there is an automatic method to evaluate the machine translation outputs, according to a metric measurement. BLUE, NIST, WER (Word Error Rate), and METEOR, are typical examples for metrics, designed to evaluate the output of machine translation.

The current study uses a group of news headlines to be translated by two main MT’s, where news headlines are considered as a ‘Block Language’, (Quirk et al., 1985, p.845). News headlines also have a special grammar, and style as stated by Swan (1996). Additionally, Iarovici and Amel (1989, p.441) define headlines as “a special kind of text, which cannot have an autonomous status”. The selected news headlines in this current study are from Arabic source language. That is, Arabic language has its unique features, which distinguishes it from other languages, Arabic has its importance and has been subjected to some experimentation in MT, especially in the US, in the very early days of MT, (Zughul & Abu-Alshaar: 2005). Izwaini (2006, p.118) states that, “*Since it was developed, Arabic machine translation has been subject to description and evaluation*” (Chalabi 2001, Farghaly & Senellart 2003, Al-Salaman 2004). Arabic has been pointed out as “notorious for complex morphology” by (McCarthy, 1985; Azmi, 1988; Beesley, 1998; Ratcliffe,1998; & Ibrahim, 2002). The view is that, Arabic as other rich morphologically languages passes through multiple stages. The translation process is difficult and represents a challenge in computational analysis technologies. These stages are called “tokenization” (Habash and Sadat, 2006). A comparative study of Arabic-English by SaeedGh (2011, p. 80) , states that, in Arabic language, each word consists of stem and vowel melody, which is equivalent to ‘al-harakaat’ in Arabic like short vowels, which are pronounced to give tone to the word that determine the meaning as proposed by McCarthy (1979, 1981) and Harris (1941). The problem is that, when word translation is accessed, it is necessary to know the words with their ‘harakaat’ or short vowels, to distinguish the form and the function of the words. Those “harakaat” are: - u (Damma) , - a (Fatha), and - i (kasra). They are used in nominative “raf” “, accusative “nasb”, and genitive “jar”, respectively (Ryding, 2005, p.30). In addition, there are two types of Arabic sentences either nominal “Jumlaismiyya” or verbal “Jumla fi’liyya”(p.58). Furthermore, Arabic language has various word orders, which includes, Subject Verb Object (SVO), Verb Subject Object (VSO), Verb Object Subject (VOS) and Object Verb Subject (OVS), which should be taken into consideration during translation process. As an illustration of Arabic into English studies, Chafia and Ali (1995) conducted a study of machine translation from Arabic into English, and from Arabic into French. Besides, Abraham and Salim (2005) have also presented algorithms to analyze Arabic into English. They argued that these algorithms have a contrasted performance compared to ” human annotation performance”.

The motivation of the study in conducting Google because Google Translation has been proven, to be “the most powerful and accurate of any of the readily available machine translation tools”(Och, 2006). In the same study, a statement implies that, the developed machine translation can be achieved “without the need to understand the

individual languages or their specific rules” (Och, 2006). On the other hand, Babylon is a computer dictionary and translation program for Microsoft Windows. The first version of Babylon was introduced in 1997. Within one year, in 1998, its number of users increased enormously and reached 4 million. Furthermore, in the year 2011, it became one of the most popular language translation applications. It can translate a full (text, Web page, and document) in 33 languages. It has a technical term, by including built-in dictionaries and community dictionaries.

Finally, translation quality is a concept which relates to the output of the translation, whether it is by a human or machine process. Linguists, philosophers and scholars are continuously discussing about the applicable criteria for good translations in order to assess their quality. This study aims to determine a better MT by comparing Google and Babylon, which would be more appropriate to be used in translating Arabic news headlines into English in terms of the Hutchins and Somers criteria (*viz. clarity, accuracy and style*).

METHOD

The study makes use of Hutchins and Somers criteria which could be summarized as follows:

The Criteria of Hutchins and Somers of Evaluation

It is important to stress that one of the main purposes of this study is derived from the role of evaluation, as to find out what machine translation systems are able and not able to do, according to the view of misunderstandings and misconceptions of transmitted message of news headlines. The evaluation is restricted on testing the raw outputs of two machine systems, specifically Google and Babylon, in reference to the manual translation that is available by the source of the data. The testing focussed on evaluating the quality of raw outputs based on the most basic principles of machine translation evaluation rather than to focus on the operations within the potential environments of systems, as it is the task of system developers. Some of these principles are: fidelity, intelligibility, and style, which they have been reflected by Hutchins and Somers (1992). The following represents the summary of these principles:

Fidelity represents the accuracy of machine translation performance. It also means to what extent that the translated output has the ‘same’ information as the original. On the other hand, intelligibility principle expresses the clarity in the translation output. In other words, it represents that the translated output should be free from obscurity, comprehensive, and understandable. The last one is style, which expresses to what extent the translation has used the language, suitable to its content and purposes.

Data of the Study

There were 40 news headlines, which were randomly chosen from three different Arabic journals, namely www.daralhayat.com, www.aljazeera.net, and www.asharqalawsat.com, dating from 1st to 30th September. The choice of these data is based on the availability of their human English translation.

Procedures of Analysis

The main procedures used in achieving the objectives of this research are stated below:

1. Collecting the data of the study which consist of Arabic news headlines with their English manual translated versions from online sources.
2. Each Arabic headline once will run into Google translator, and then into Babylon translator, to be translated into English.
3. The outputs of both Google and Babylon will be listed in one table.
4. To fulfil the evaluation objective, the researcher had distributed a questionnaire to a group of evaluators. The distributed questionnaire was based on the criteria provided by Hutchins and Somers (1992). The group of evaluators consists of 28 professionals whose native language is Arabic, and who work in different Iraqi Universities, and have good English Language proficiency.

The Evaluators Assessment

This part is the most important process, which is to calculate the human judgments based on the assigned questionnaire. The current study conducted 40 machine-translations of Arabic news headlines into English. The evaluators were asked to consider each Arabic headline and its machine-translated outputs to examine the three parameters which are provided in the questionnaire. The parameters consisted of three criteria: *Clarity, Accuracy, and Style*. Each criterion is defined according to Hutchins and Somers (1992). For each criterion there were 4 scores. There were 28 evaluators who participated in the assigned questionnaire. The average of each output was calculated based on the following statistical equation:

$$Av = \frac{X_1 + X_2 + X_3 \dots + X_{28}}{n\{evaluator\}}$$

Av = average
X = the score of the evaluator
n = the number of the evaluators

Then, by summing up the averages of each outputs of the same parameter and dividing them by the number of outputs, we obtained the total average of each parameter according to the following equation:

$$Total\ Av = \frac{Av_1 + Av_2 + \dots + Av_{40}}{n\{output\}}$$

For example, to find the average of the clarity criterion of translated output for Headline (1) by Google under the aspect of clarity: *How easily can you understand the translation?*

1 – Not understandable	1 participant/28participants= 3.6 %
2 – Only small part understandable	0 participant/28 participants = 0.0 %
3 – Mostly understandable	5 participants /28 participants =17.8 %
4 – Fully understandable	22 participants /28participants = 78.6 %

As shown above, the first answer, “Not understandable”, was chosen only by one out of 28 participants, giving a score of 3.6%. However, no participant chose the second answer, “Only small part understandable”, and as a result the score was 0%. In contrast, the third answer, “Mostly understandable”, was selected by 5 participants out of the total of 28 evaluators. However, the fourth answer had the highest score of 78.6%, as it was chosen by 22 participants. Consider the following Figure 1 which illustrates what mentioned above:

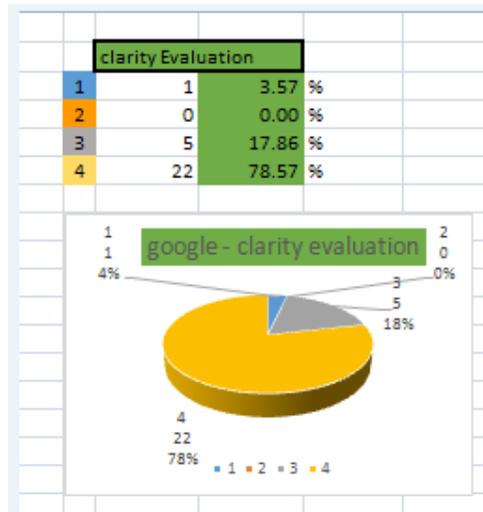


Figure 1. Percentage of the participants answers

Then, the Average will be calculated as the following :

$$Av = \frac{4 + 4 + 4 + 3 + 4 + 1 + 4 + 4 + 4 + 4 + 3 + 3 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 3 + 4 + 4 + 4 + 3 + 4}{28} = \frac{104}{28} = 3.7$$

The following Table 1 shows the process of Google output of Headline (1). Consider the part related to the clarity criterion as shown in Figure 2.

Table 1. Google translation output for Headline 1

Arabic Headline: n .1	هنية يدعو الرئيس المصري للتجديد بأقامة المنطقة التجارية الحرة مع غزة
Google translation	Haniyeh calls for Egyptian President to accelerate the establishment of free trade zone with Gaza

participant	Google translation (GT) Evaluation		
	clarity Evaluation	Accurac Evaluation	style Evaluation
1	4	4	4
2	4	4	3
3	4	3	3
4	3	4	4
5	4	4	3
6	1	2	2
7	4	4	4
8	4	4	4
9	4	4	3
10	4	4	3
11	3	3	3
12	3	3	3
13	4	4	4
14	4	4	4
15	4	4	4
16	4	4	4
17	4	3	3
18	4	4	3
19	4	4	3
20	4	4	3
21	4	4	3
22	4	4	4
23	3	2	3
24	4	4	4
25	4	4	3
26	4	4	4
27	3	3	4
28	4	4	4
	104	103	91

Figure 2. The participants answers of parameters

The same process was carried out to determine the clarity of the Babylon output for headline (1):

- 1 – Not understandable 0 participant /28 participants = 0.0 %
- 2 – Only small part understandable 1 participant /28 participants = 3.6 %
- 3 – Mostly understandable 5 participants /28 participants =17.8 %
- 4 – Fully understandable 22 participants /28 participants = 78.6 %

As shown above, the answer, “not understandable”, scored 0.0% as no participant chose this answer, while one participant selected the second answer, “Only small part understandable”, giving a score of 3.6%.The third answer, “Mostly understandable”, obtained a score of 17.8%, as it was chosen by 5 out of 28 participants. On the other hand, the fourth answer, “Fully understandable”, was selected by 22 evaluators, giving an average score of 78.6% (Refer to Figure 3).

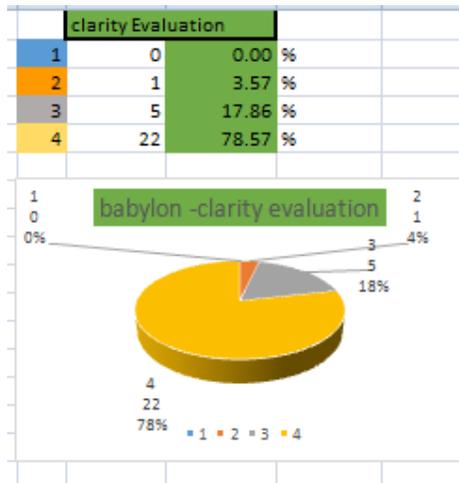


Figure 3. Percentage of participants answers

Moving to the Average, consider the following :

$$\begin{aligned}
 Av &= \frac{4 + 4 + 4 + 4 + 3 + 4 + 2 + 4 + 4 + 4 + 4 + 4 + 3 + 3 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 3 + 4 + 4 + 4 + 3 + 4}{28} \\
 &= \frac{105}{28} \\
 &= 3.8
 \end{aligned}$$

The following equation is used to find the percentage for each parameter or criterion:

$$\frac{\text{Total (Av)} \times 100}{4}$$

Consider the following Table 2 which shows the output of the same Headline produced by Babylon and the representative averages for each parameter is shown in Figure 4.

Table 2. Babylon translation output for Headline 1

Arabic Headline n. 1	
هنية يدعو الرئيس المصري للتجويل بإنقمة المنطقة التجارية الحرة مع غزة	
Babylon translation	
Haniyya calls for Egyptian president to accelerate the establishment of free trade zone with Gaza	

participant	Babylon translation (BT) Evaluation		
	clarity Evaluation	Accurac Evaluation	style Evaluation
1	4	4	4
2	4	4	4
3	4	3	3
4	3	4	4
5	4	4	3
6	2	1	1
7	4	4	4
8	4	3	3
9	4	4	3
10	4	4	4
11	3	3	3
12	3	3	3
13	4	4	4
14	4	4	4
15	4	4	3
16	4	4	4
17	4	3	3
18	4	4	2
19	4	4	2
20	4	4	2
21	4	4	2
22	4	4	4
23	3	2	3
24	4	4	4
25	4	4	3
26	4	4	4
27	3	3	3
28	4	4	4
	105	101	90

Figure 4. The participants answers of parameters

FINDINGS

The following sections will show the results of each criterion for each system. The results are based on the evaluators' assessment of the provided questionnaire, as well as the results of the preferred system in translating such data. The overall calculated averages of participants' responses for parameters for all headlines is shown in Figure 5 comparing Google and Babylon.

Headline s. nr.	Google translation (GT) Evaluation			Babylon translation (BT) Evaluation		
	clarity Evaluation	Accuracy Evaluation	style Evaluation	average clarity Evaluation	average Accuracy Evaluation	style Evaluation
1	3.7	3.7	3.3	3.8	3.6	3.2
2	3.7	3.7	3.3	3.8	3.6	3.2
3	3.7	3.7	3.3	3.8	3.6	3.2
4	3.7	3.7	3.3	3.8	3.6	3.2
5	3.7	3.7	3.3	3.8	3.6	3.2
6	3.7	3.7	3.3	3.8	3.6	3.2
7	3.7	3.7	3.3	3.8	3.6	3.2
8	3.7	3.7	3.3	3.8	3.6	3.2
9	3.7	3.7	3.3	3.8	3.6	3.2
10	3.7	3.7	3.3	3.8	3.6	3.2
11	3.7	3.7	3.3	3.8	3.6	3.2
12	3.7	3.7	3.3	3.8	3.6	3.2
13	3.7	3.7	3.3	3.8	3.6	3.2
14	3.7	3.7	3.3	3.8	3.6	3.2
15	3.7	3.7	3.3	3.8	3.6	3.2
16	3.7	3.7	3.3	3.8	3.6	3.2
17	3.7	3.7	3.3	3.8	3.6	3.2
18	3.7	3.7	3.3	3.8	3.6	3.2
19	3.7	3.7	3.3	3.8	3.6	3.2
20	3.7	3.7	3.3	3.8	3.6	3.2
21	3.7	3.7	3.3	3.8	3.6	3.2
22	3.7	3.7	3.3	3.8	3.6	3.2
23	3.7	3.7	3.3	3.8	3.6	3.2
24	3.7	3.7	3.3	3.8	3.6	3.2
25	3.7	3.7	3.3	3.8	3.6	3.2
26	3.7	3.7	3.3	3.8	3.6	3.2
27	3.7	3.7	3.3	3.8	3.6	3.2
28	3.7	3.7	3.3	3.8	3.6	3.2
29	3.7	3.7	3.3	3.8	3.6	3.2
30	3.7	3.7	3.3	3.8	3.6	3.2
31	3.7	3.7	3.3	3.8	3.6	3.2
32	3.7	3.7	3.3	3.8	3.6	3.2
33	3.7	3.7	3.3	3.8	3.6	3.2
34	3.7	3.7	3.3	3.8	3.6	3.2
35	3.7	3.7	3.3	3.8	3.6	3.2
36	3.7	3.7	3.3	3.8	3.6	3.2
37	3.7	3.7	3.3	3.8	3.6	3.2
38	3.7	3.7	3.3	3.8	3.6	3.2
39	3.7	3.7	3.3	3.8	3.6	3.2
40	3.7	3.7	3.3	3.8	3.6	3.2
total	148.6	147.1	130.0	150.0	144.3	128.6
average	3.7	3.7	3.3	3.8	3.6	3.2

Figure 5. The calculated averages comparing Google and Babylon

Clarity:

Based on Figure 5, *Clarity* was the first parameter in which the participants were asked to evaluate. There were only minimal differences between the clarity of the Google and Babylon translations for each of the forty (40) outputs of headlines. From Figure 6, it is obviously shown that both the two translators were graded with an average of 3.2 out of the highest value of 4. We can say that the evaluators assessed both the Google and Babylon outputs as being equally understandable. The score was closest to 3, which indicates that “Mostly understandable” was the answer to the question “How easily can you understand the translation?”. Accordingly, the evaluators’ estimation for both Google and Babylon was 80% *clarity*.

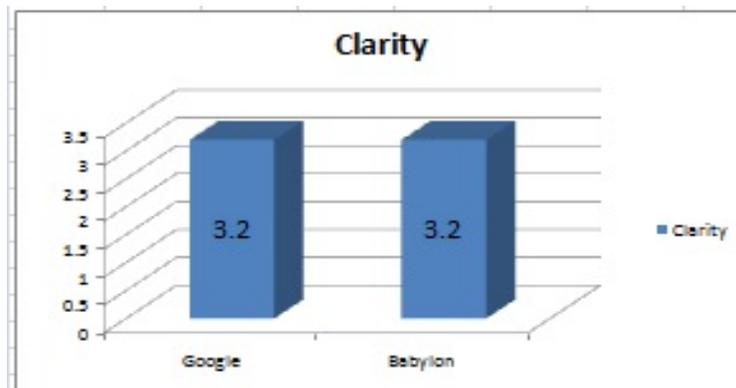


Figure 6. Clarity

Accuracy

The second parameter to be marked by the evaluators was *accuracy*. Referring to Figure 7, overall, Google scored higher than Babylon in terms of *accuracy*. Out of the highest value of 4, Google had an average score of 3.1, whereas the combined average score of Babylon was 3.0. The assessment of the criteria indicated that both Google and Babylon were closest to the score of 3, which gave the evaluators' answer to the question, "**To what extent does the translation contain the 'same' information as the source text?**" It was clear that these two averages illustrated that there was a significant variation between Google and Babylon, as shown by the following rating: 77.5% for Google and 75% for Babylon. Accordingly, Google was highly regarded by the evaluators to be more accurate than Babylon, as can be seen in the following Figure:

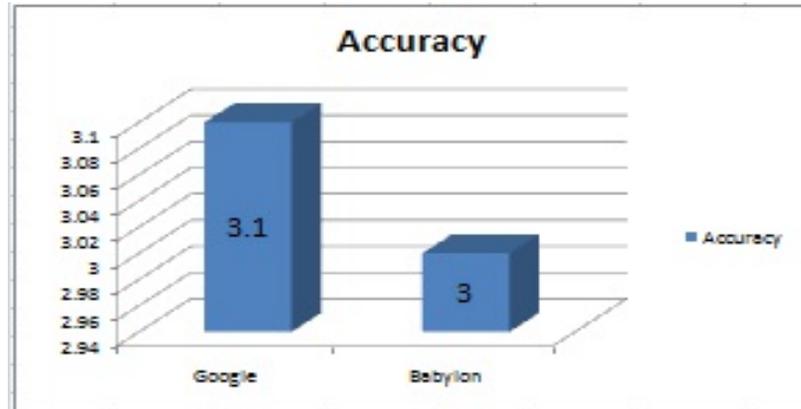


Figure 7. Accuracy

Style

The third parameter which the evaluators were asked to score was *style*. Babylon scored higher than Google, where the average for Babylon's average was 2.9 out of 4, which represented the highest rating. Google's average meanwhile was 2.8. Hence, the average of Google's style was considered as the lowest average out of the three criteria. It was apparently shown by accounting the percentage of each style average that the evaluators found that the style of the Babylon outputs was better than the style of the Google outputs. Thus, Google had 70% and Babylon had 72.5% of *style*. Concerning the criteria, the evaluation was based on answering the following question: "**Is the language used appropriate for a software product user manual? Does it sound natural and idiomatic?**" The answer revealed that Babylon somehow produced a more acceptable *style* in its outputs than the *style* of the Google outputs, as shown in the following Figure 8.

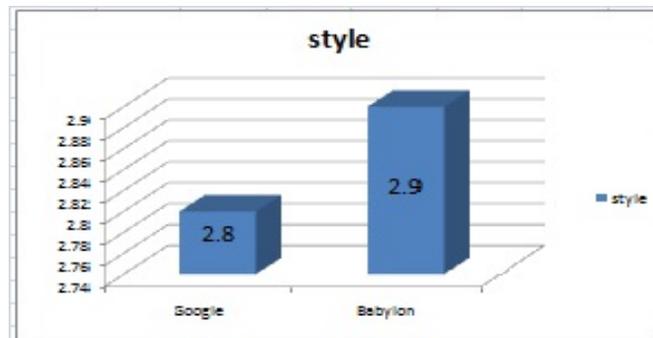


Figure 8. Style

Last but not least, the evaluators were asked to give their preferred system for translation. Interestingly, the results showed that 16 out of the 28 chose to use Babylon while the remaining 12 preferred Google. The following Figure 9 illustrates the percentage obtained by each system. 43% chose Google, while 57% preferred Babylon. . Consider the following Figure:

	A	B	C	D	E
3					
4					
5	participant	GT		BT	
6	1			1	
7	2			1	
8	3			1	
9	4	1			
10	5			1	
11	6	1			
12	7			1	
13	8			1	
14	9			1	
15	10	1			
16	11	1			
17	12	1			
18	13			1	
19	14	1			
20	15			1	
21	16	1			
22	17			1	
23	18			1	
24	19	1			
25	20			1	
26	21			1	
27	22	1			
28	23	1			
29	24			1	
30	25	1			
31	26			1	
32	27			1	
33	28	1			
34	TOTAL	12		16	
35	average	43 %		57 %	

Figure 9. The participants preferred system

To illustrate the percentage of their choice, see the following Figure 10.

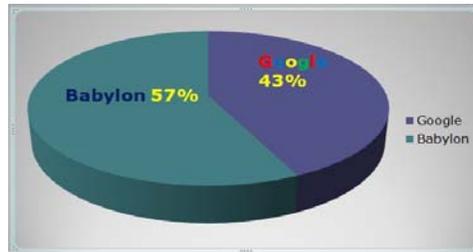


Figure 10. The preferred system

Finally, the evaluators’ assessment indicated that the selected machine translators had *clarity*, *accuracy*, and *style* but each had different values. They also revealed that the majority of the evaluators preferred to use the outputs from Babylon rather than from Google.

For the third objective, the results showed that the evaluators’ estimation was different for each system according to the provided criteria which they had to examine. Both systems had the same degree of value only in the criterion of clarity, whereas each system scored different values for the other two criteria of accuracy and style. For accuracy, Google got a higher value than the Babylon system. However, Babylon got a higher score in terms of style. The following Figure 11 shows the average values of the systems:

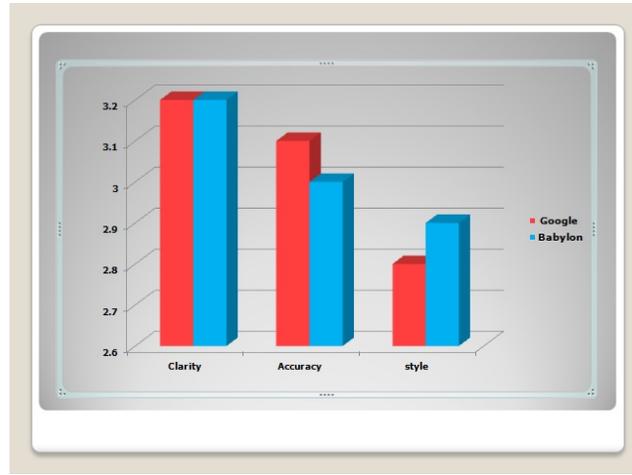


Figure 11. The averages of each system

In the above Figure 11, the score for the Google and Babylon systems was 3.2 in terms of clarity, while Google got an average of 3.1 and Babylon got an average of 3.0 in terms of accuracy. However, Babylon got a 2.9 average for style, which is higher than Google's average of 2.8. The following Figure (2.8) shows the percentage of each system with regard to these averages.

The results of the assigned questionnaire show that the evaluators preferred to use Babylon than Google. The former scored 57% of evaluators' preference, while 43% preferred that the latter be used in translating such data. The results also demonstrate that both translators, Google and Babylon, had the same score of 80% for Clarity, in contrast to the second parameter, 'Accuracy,' for which Google scored a higher value than Babylon. The former scored 77.5%, whereas the latter scored 75%. However, Babylon had a higher value of 72.5% for Style, in contrast to Google's score of only 70%. In this case, Babylon focused on 'Style' more than Google from the evaluators' point of view.

IMPLICATIONS AND CONCLUSIONS

Online MT can be used for the purpose of learning from school to tertiary level because it has the characteristics of educational technologies that can help students, especially for students who want to pursue a foreign language. MT is commonly used to understand a second language text and express their ideas. MT has been shown to accelerate the translation work and very time saving. MT use in translation actually shortens some steps as used in the human translation. One no longer need to search for words, flipping page after page which is certainly time consuming then write back. Instead, the software can easily translate the content and quality translation results with word choices. In the era of globalization, the dominance of such information is a value added for individuals and the organization. Information can be obtained from a variety of languages throughout the world. With the availability of MT, such information can be obtained easily and cost effective without high investment. On the other hand, if a translation done by a professional translator, translation based on a per page basis would certainly be very costly and compared to the use of MT which involves a very minimal cost.

Confidentiality is also one of the characteristics found in the nature of MT-aided translation. MT usage ensures information translated is protected whereas; the submission of documents which holds sensitive information may risk leakage if given to a human translator. The software in MT has been designed for use in universal fields. MT is very suitable for use in science, literature, language and linguistics, and others whereas; human translation only covers specific areas of expertise.

Undoubtedly, MT has many benefits that can help students transfer information into preferred language. It is necessary for them to be more cautious when doing translation work since there are areas that cannot be translated as cultural aspects associated with the accuracy of meaning which cannot be produced by machine translation consistently. One can only obtain information in the form or essence of the draft document and it is not necessarily fully accurate. This is because MT is only capable of conducting literal translation of the words without understanding the actual information in context that may need to be corrected manually later. Another flip side of MT is that it cannot handle ambiguities that exist because it was created under the laws of systematic and formal rules of the language and certainly could not translate words based on experience, emotions, values, and mental outlook compared to human translation. However, online machine translation systems are

continuously undergoing development, and the outputs might be improved in the near future to help students' learning more effectively.

REFERENCES

- Alsaman, S. (2004). The Effectiveness of Machine Translation. In IJAIS, International Journal of Arabic-English Studies, 5, pp.145-160
- Abraham, I., Salim, R. (2005). A Maximum Entropy Word Aligner for Arabic-English Machine Translation. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 89–96, Vancouver.
- ALPAC (1966). Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, 1966. (Publication 1416.)
- Chalabi, A. (2001). Sakhr Web-based Arabic/English MT Engine. <www.elsnet.org/arabic2001/chalabi.pdf> [10 August 2010].
- Chafia, M., Ali, Mili. (1995). Machine Translation from Arabic to English and French Information Sciences. Elsevier Science Inc. 655 Avenue of the Americas, New York, NY 10010, 3(2):91-109.
- Cyrus L. (2009). Old Concepts, New Ideas: Approaches to Translation Shifts. University of Münster.
- Doyon, J., Taylor, K. B., and White, J. S. (1998). "The DARPA Machine Translation Evaluation Methodology: Past and Present." Philadelphia: AMTA98.
- Farghaly, A., Senellart, J. (2003). Intuitive Coding of the Arabic Lexicon. A paper read at The MT Summit IX – Workshop: Machine translation for semitic languages, New Orleans, USA.
- Flanagan, M. (1994). Error Classification for MT Evaluation. In Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas, Columbia, MD.
- Groves, D., Way, A. (2006). Hybrid data-driven models of machine translation. Springer Science & Business Media B.V. Volume 19 PP 301-323.
- Harris, Z. (1941). Linguistic Structure of Hebrew. Journal of the American Oriental Society.
- Hutchins, W.J. (1986). Machine Translation. Past, Present, Future. West Sussex, England: Ellis Horwood Limited.
- Hutchins, J. and Somers, H. (1992). An Introduction to Machine Translation. London: Academic Press Limited
- Hovy, Eduard H., Margaret King and Andrei Popescu-Belis. (2002), 'Computer-Aided Specification of Quality Models for MT Evaluation', in LREC 2002: Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain, pp. 1239–1246.
- Hovy, Eduard H., Margaret King and Andrei Popescu-Belis: (2002), 'Principles of Context-Based Machine Translation'. Machine translation 17: 43-75, 2002. © 2003 Kluwer Academic Publishers. Printed in the Netherlands.
- Iarovici, E. and Amel, R. (1989). The Strategy of the Headline. Semiotica. 77(4): 441- 459.
- Izwaini, S. (2006). Problems of Arabic Machine Translation: Evaluation of Three Systems. Proceedings of the International Conference at the British Computer Society (BSC), London.
- Marcu, D. (2001). Towards a unified approach to memory- and statistical-based machine translation. In: Association for Computational Linguistics: 39th annual meeting and 10th conference of the European chapter, Toulouse, France, pp. 378–385.
- McCarthy, J. (1981). A Prosodic Theory of Non concatenative Morphology. Linguistic Inquiry. Eport.
- Och, F. J. (2006). The Alignment Template Approach to Statistical Machine Translation. MIT Press Journals, 30(4), 417-449.
- Papineni, K., Salim, R., Todd, W., Wei-Jing, Z. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318.
- Quirk, R., et al. (1985). A Comprehensive Grammar of the English Language. Longman. London and New York. © Longman Group Limited 1985.
- Richardson, S., Dolan, W., Menezes, A., Pinkham, J. (2001). Achieving commercial-quality translation with example-based methods. In Proceedings of MT Summit VIII, Santiago De Compostela, Spain.
- Ryding, Karin C. (2005). A Reference Grammar of Modern Standard Arabic.
- Salem, Y., Arnold, H., Brian, N. (2008). Implementing Arabicto- English Machine Translation using the Role and Reference Grammar Linguistic Model, in Proceedings of the Eighth Annual International Conference on Information Technology and Telecommunication (ITT 2008), Galway, Ireland. (Runner-up for Best Paper Award).
- Schiaffino R. & Zearo F. (2005). 46th ATA Conference, Seattle. Aliquantum and L10NBRIDGE.

- Swan M. (1996). Practical English Usage. International Student's Edition. Second Edition .Oxford University Press. P: 359- 369
- Tahir, G.R., Asghar, S., Masood, N. (2010). Knowledge Based Machine Translation. Proceedings of International Conference on Information and Emerging Technologies (ICIET), Karachi, Pakistan pp. 1-5.
- White, J., O'Connell, T. and O'Mara, F. (1994) "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches". Proceedings of the 1st Conference of the Association.
- Williams, M (2009). Translation Quality Assessment. University of Ottawa. Vol 2, No 1. 2009. pp. 3 – 23.
- Zughoul, M, Abu-Alshaar, A. (2005).s: A Historical Perspective Translators' Journal 50 (3): 1022–1041. Retrieved from <http://id.erudit.org/iderudit/011612ar>. Retrieved 2 June 2011.