Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English

Zakaryia Mustafa Almahasees

***Abstract*—Online machine translation (OMT) systems are widely used throughout the world freely or at low cost. Most of these systems use statistical machine translation (SMT) that is based on a corpus full with translation examples to learn from them how to translate correctly. Online automatic machine translation systems differ widely in their effectiveness and accuracy. Therefore, the wide spread of such translation platforms make it necessary to evaluate the output in order to shed light on the capacity and usability of each system. The present study have selected the most prominent translation systems, Google and Microsoft to test which system is better and more reliable in rendering English<>Arabic translation. To conduct the study, the researcher has chosen automatic evaluation of the two system outputs by using the most popular automatic evacuation metric BLEU. The study's corpus consists of 25 Arabic sentences extracted from Petra News Agency of Jordan with its human reference translation from the English version of Petra. The result of the research showed that Google translate achieves better results than Microsoft Bing in comparison to human referenced translation. However, Machine Translation (MT) is still far from reaching fully automatic translation of a quality obtained by human translators.**

***Index Terms*—Automatic evaluation, Arabic machine translation (AMT), Arabic MT evaluation, Google evaluation, Microsoft Bing evaluation, BLEU.**

I. Introduction

Machine Translation (MT) is the use of computers in the process of translation from one natural language into another. The combination of translation and technology has resulted in recent years in automatic translation or translation by computer called Machine Translation, which allow computers to translate from one language into another. Theoretically, it is considered a branch of computational linguistics. Studying Machine Translation covers three interrelated areas of knowledge: translation, linguistics and translations' software.

Besides, most of scientific research and knowledge contribution are conducted in English, so researchers, students and people in general who do not master English use MT as a guide to look up for the meaning of certain words or translate whole texts, MT systems offering whole text translation, and providing free and quick service. What is the degree of accuracy of MT? How close is MT translation to human translation? The paper sheds some light on the usability and capacity of Google and Microsoft Bing in rendering Arabic political sentences into English. To do so, the paper evaluates MT's Output in comparison to human referenced translation that is available for each chosen extracted sentence from Petra News [1].

Evaluation of MT is necessary and used mostly by systems developers and researchers to verify the effectiveness of MT. Evaluation of MT can be done by developers, researchers, and translators. At the same time, there are different methods used to evaluate MT efficiency. Some of them are human evaluation of MT which is based usually on face to face interviews, surveys, or bilingual speakers to rate the accuracy of translation. These methods are expensive, and at times inconsistent due to different views on the best translation. On the other hand, automatic evaluation measures the ability of the system to provide appropriate translation. Thus, the study uses a well-known automatic evaluation metrics to measure the closeness of MT to human professional translation, BLEU (Bilingual Evaluation Method) introduced by [2]. It is widely used to evaluate the closeness of MT to human translation, and it has proven its effectiveness in evaluating Machine translation systems. Reference [3] shows that "BLEU is based on a core idea to determine the quality of any machine translation system which is summarized by the closeness of the candidate output of the machine translation system to reference (professional human) translation of the same text." Therefore, the study has chosen BLEU due to its consistent and stable results. The next section of this paper gives a review of related literature on the topic. It is followed by a section three presents the methodology of the study. Section 4 presents a discussion of the two translation systems and how BLEU evaluates the output of the chosen systems; while section 5 sums up the study that based on my main findings and provide recommendations for future research.

II. Review of Relevant Literature

Several studies have been conducted to evaluate the quality and accuracy of machine translation by using automatic evaluation metrics such as BLEU [2], Meteor, NIST, F-Function and the like. Reference [4] indicates in *An Evaluation Tool for Machine Translation* presented the typical method to establish an effective tool to evaluate the accuracy of various MT systems. They discussed two important criteria for the quality of MT Systems' output: WER (Word Error Rate) and SSER (Subjective Sentence Error Rate). They identify WER and SSER measures as

Manuscript received October 10, 2016; revised March 18, 2017. This Research is funded by the University of Western Australia, Graduate Research Grants (Research Travel Grant).

Zakaryia Almahasees is with University of Western Australia, Australia (e-mail: zakaryia.almahasees@research.uwa.edu.au).

doi: 10.18178/ijlll.2017.3.1.100

yielding fast, semiautomatic, consistent and suitable results. Moreover, Reference [5] contends in *Precision and Recall of Machine Translation* that Precision, Recall and F-Measure are more reliable than BLEU metric of [2]. They add that these metrics are usually used for information retrieval, data mining and search engines. Reference [6] in *User-Centered Evaluation for Machine Translation of spoken Language* introduced User-Centered method to evaluate MT, which is based on comparing MT output to human referenced translation. The evaluation of MT on Arabic to English and Mandarin to English ranks MT output in comparison to referenced human translation by an expert to judge which output is closer to human translation.

Reference [7] in *Using Multiple Edit Distances to Automatically Grade Outputs from Machine Translation Systems* presented an evaluation method that is a subsystem of SSMT (Speech–to-speech MT systems. The method is "Grader based on Edit Distance" that compute the score of MT output by using a decision tree. They conducted several experiments, and they claimed that Red (Radar Based Edit Distance) is more accurate than BLEU. BLEU evaluation method is language independent and can be used to assess any natural language. Reference [8] shows that conducted in *Extending BLEU Evaluation Method with Linguistics Weight* research to improve the effectiveness of BLEU method, and they succeeded through the use of multiple Ngram weights. Reference [9] evaluates 2-way Iraqi Arabic–English speech translation systems using automated metrics. They found that automatic translation of Iraqi Arabic correlates with human judgements. Moreover, Reference [10] evaluated Arabic machine Translation by using three automatic measures: BLEU, F1 and F mean. Their evaluation is based on Universal Networking Language (UNL) and the Interlingua approach for machine translation. Reference [11] in *Arabic Machine Translation Survey* discussed the challenges for Arabic Machine Translation. The research found that it is difficult to find a suitable machine translation that meet human requirements.

The previous research highlights the importance of automatic evaluation. Several scholars contributed to the field by using a set of metrics that measure the closeness of MT output to reference range from ranking metrics to grader systems, BLEU, and human participants in the process of MT evaluation. BLUE, is "the best known and best adopted machine evaluation for machine translation" as stated in EuroMatrix, [12] Moreover, Reference [13] states that BLUE is used to "determine the quality of any machine translation system which is summarized by the closeness of the candidate output of the machine translation system to reference (professional human) translation of same text." Reference [2] confirms that BLUE use n-gram precision to differentiate between strong and weak translation of MT. Reference [14] confirms the importance of comparing the output of MT to existing translation to identify their strengths and limitations. They state "in order to find the errors in a translation, it is useful to have one or more reference translation in order to contrast the output of MT system with a correct text." Thus, BLEU is the most popular due its best correlation with human judgments. This is the reason why the present study adopts BLEU.

III. Methodology

The current study adopts BLEU method to evaluate Google and Microsoft Bing outputs to verify the effective and the best translation system in translating political texts. The corpus of the study has selected from Petra News Agency journalistic reports [1] on local, global and regional issues. The source sentence, Arabic, is inputted to two selected translation systems, and the output of the MT is English. Then, MT output is compared to referenced human translation that is available at English version of Petra News Agency website.

The source sentence is extracted from Petra News Agency with the reference human translation that is available at English version of Petra News Agency. The researcher analyzed the sentences to find the Ngram strings to calculate the precision of each sentence in comparison to human referenced translation. In this respect, the IBM formula of BLEU [2] is adopted to measure the precision as follow:

1) Brevity Penalty: BP

$$= \begin{cases} 1 & if c > r \\ e(1 - r/c) & if c \leq r \end{cases}$$

2) $$BLUE = BP \times exp\left(\sum_{n=1}^{N} wn \log pn \right)$$

IV. Discussion

The present section presents the analysis and measurement of the MT output to conduct this study to verify the effectiveness of MT output through text similarity metric, BLEU. To explain how BLEU works, the following example elucidates how to measure the closeness between the candidate (MT output) and human referenced translation. Our example is:

Source Text (Arabic):

The selected sentence is inputted to Google and Microsoft Bing systems and the output is as follows:

عاد جلالة الملك عبد الله الثاني الى أرض الوطن, امس الجمعة بعد ان اختتم زيارة عمل الى مدينة نيويورك.

Google: *His majesty King Abdullah II returned home, on Friday, having concluded a working visit to New York City.*

Bing: *His majesty King Abdullah II returned home on Friday, after concluding his visit to New York City.*

Reference Translation is taken from the English version of Petra News Agency of Jordan.

Reference: *His majesty King Abdullah II returned home on Friday, following a working visit to New York.*

Initially, the analysis of the MT output is to find N-gram strings: Unigram (one word), Bigram (two words), trigram (three words) and tetra gram (four words). The similarity in number of unique words is complemented by a test on a similarity in strings of 2, 3 or 4 words; therefore the closer the words and their ordering is to the reference translation, the better the precision score. The results are shown in Table I and Table II.

TABLE I: PRECISION VALUES

MT system / Ngram	Google	Bing
Unigram	16/21	14/19
Bigram	10/20	8/18
Trigram	7/19	6/17
Tetragram	5/18	4/16

TABLE II: NGRAM PRECISION VALUE

MT System / Ngram p	Google	Bing
Precision (1)	0.76	0.73
Precision (2)	0.50	0.44
Precision (3)	0.36	0.35
Precision (4)	0.27	0.25

Secondly, we calculated the score of BLEU by computing Brevity Penalty value by choosing the best system.

$$\text{Brevity Penalty BP} = \begin{cases} 1 & if c > r \\ e(1 - r/c) & if c \leq r \end{cases}$$

C= MT output length r = Reference translation length

Brevity Penalty rule is that if the candidate is longer than the reference is then BP is set at (1), while if it is shorter, the result is (0). In our example for Google, C=21, R=16, and when 21<16 then BP= 1 and for Bing C= 19, r=16 and when 19<16 then also BP=1. Then, we use the Brevity Penalty (BP) results from the first equation, to compute the final BLEU Score. Reference [13] indicates that BP "n-gram precision penalizes candidate sentences found shorter than their reference counter parts, also it penalize candidate sentences which have over generated correct word forms." Reference [2] mentions that BLEU metric precision value ranges from 0 to 1, where the translation that has a score of 1 is identical to a reference translation.

$$(2) BLEU = BP \times exp\left(\sum_{n=1}^{N} wn \log pn\right)$$

BLEU= *exp* (1/4 ×log(16/21)+1/4 ×log(10/20)+1/4 ×log(7/19)+1/4 ×log(5/18)= 0.70

The result of computing BLEU score of Google is 0.70, while for Microsoft Bing is as follow:

Brevity Penalty for Bing = 1

BLEU= *exp*(1/4 ×log(14/19)+1/4 ×log(8/18)+1/4 ×log(6/17)+1/4 ×log(4/16)= 0.68

The analysis showed that BLEU score for Google is 0.70, and the BLEU score for Microsoft Bing is 0.68 as shown in Fig. 1. This result shows that Google translate achieves better results than Microsoft Bing Translation system because higher BLEU score for any machine translator means that it is better than its lower counterpart. It is indicated through the results that Google is more efficient and better than Microsoft Bing in translating Arabic political sentences into English. However, the two MT systems are equal sometimes, and they are still far away from achieving overall quality with reference to human translation.

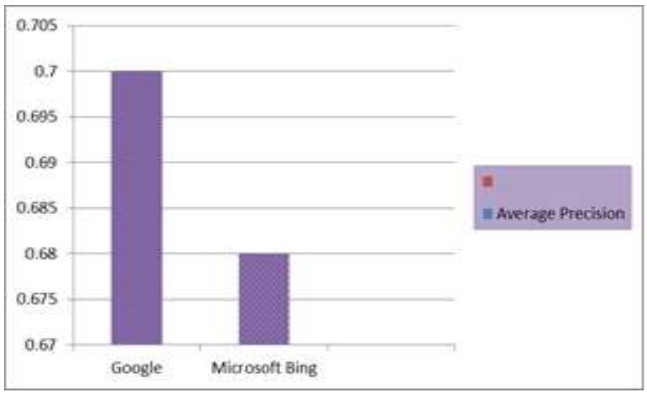

Fig. 1. BLEU score for Google and Microsoft Bing.

As whole the average results of Google translate and Microsoft Bing show that Google has shown in the table of results of the corpus of political registers. The next table shows the precision ratio to the same metrics on one sentence taken from arrange of texts.

TABLE III: PRECISION VALUE FOR EACH SENTENCE OF THE SELECTED REGISTERS

Register / MT	Global Issues	Regional /local Issues	Political Relations	International Commitments
Google	0.51	0.50 0.35 0.35 0.40 0.48 0.70	0.69	0.64
Bing	0.25	0.55 0.34 0.64 0.39 0.43 0.68	0.53	0.68

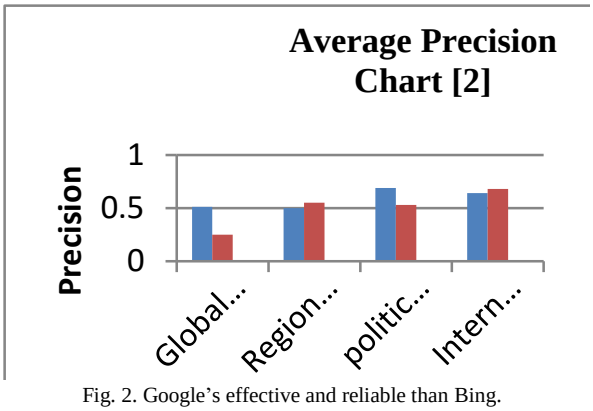

Fig. 2. Google's effective and reliable than Bing.

The above Table III and Fig. 2 show that Google has proven its effective and reliable than Bing in rendering different sentences that represent different political registers including global, regional, political relations and international commitment. However, there is a slight similarity between Google and Bing, where Google has the greatest text similarity to reference which give google a higher result in comparison to Bing, where some Bing translation lack the sequences on trigram and tetra gram sequences.

V. Conclusion

Translation from Arabic to English is a challenging task due to differences in syntactical, morphological and semantic features between the two languages. The study evaluated the translation of Arabic to English political sentences that is taken from journalistic reports of Petra News Agency. The study adopted BLEU metric to find out values of MT precision in comparison to human referenced translation. The study found that Google shows higher correlation to human referenced translation and it is more effective and better than Microsoft Bing. However, Machine Translation (MT) is still far from reaching fully automatic translation of a quality obtained by human translators.

VI. Recommendations

The study recommends future research in correlation of Automatic and manual evaluation methods, which would provide strong and more reliable results, regards the usability of MT systems. Moreover, it is recommended an analysis of MT output from a linguistics point of view to provide the Machine translation field with a rich literature which help the systems designers to focus on the lagging behind points.

Acknowledgment:

My profoundest gratitude is due to my Supervisor Prof. Dr. Helene Jaccomard for her extremely intellectual and generosity, which help me to broaden my understanding of MT field thematically and systematically.

References

[1] Petra News Agency. (2017). [Online]. Available: http://www.petra.gov.jo/Public/Main_arabic.aspx?lang=1&site_id=2

[2] K. Papineni *et al.*, "BLEU: A method for automatic evaluation of machine translation," in *Proc. the 40th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, 2001, pp. 311-18.

[3] Al-Kabi *et al.*, "Evaluating English to Arabic machine translation using BLEU," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, 2013.

[4] S. Nie&en *et al.*, "An evaluation tool for machine translation: Fast evaluation for MT research," in *Proc. the 2nd International Conference on Language Resources and Evaluation*, 2000, pp. 39-45.

[5] Melamed, *et al.*, "Precision and recall of machine translation," in *Proc. the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL*, 2003, vol. 2, pp. 61-63, Association for Computational Linguistics.

[6] D. D. Palmer, "User-centered evaluation for machine translation of spoken language," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, vol. 5, pp. v-1013, IEEE.

[7] Y. Akiba *et al.*, "Using multiple edit distances to automatically grade outputs from Machine translation systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 393-402, 2006.

[8] M. Yang *et al.*, "Extending BLEU evaluation method with linguistic weight," in *Proc. the 9th International Conference for Young Computer Scientists*, 2008, pp. 1683-1688, IEEE.

[9] S. Condon *et al*. (2012). "Evaluation of two Iraqi Arabic–English speech translation systems using automated metrics," *Machine Translation*, vol. 1-2, pp. 159-176, 2012.

[10] N. Adly and S. Ansary, "Evaluation of Arabic machine translation system based on the universal networking language," in *Proc. the 14th International Conference on Applications of Natural Language to Information Systems*, 2009, pp. 243-257.

[11] A. Alqudsi *et al.*, "Arabic machine translation: A survey," *Artificial Intelligence Review*, vol. 42, no. 4, pp. 549-572, 2014.

[12] EuroMatrix. (2007). 1.3: *Survey of Machine Translation Evaluation. Satistical and Hybird Machine Translation between all European Languages*. [Online]. Available: http://www.euromatrix.net/deliverables/Euromatrix_D1.3_Revised.pdf

[13] Al-Kabi *et al.*, "Evaluating English to Arabic machine translation using BLEU," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, 2013.

[14] D. Vilar *et al.*, "Error analysis of statistical Machine Translation output," in *Proc. LREC*, 2006, pp. 697-702.

Zakaryia Mustafa Slameh Almahasees was born in Jordan on November 14, 1986. Currently, he is a PhD student in translation at University of Western Australia. His main research is translation and technology: machine translation. He is working on English-Arabic machine translation systems. He worked as full-time lecture for three years in English literature and translation. He got his PhD scholarship from Applied Science University, Jordan in 2015. He completed his MA in English language and literature in 2012, Jordan and his BA in English in 2008 from Yarmouk University, Jordan.