AWEJ for Translation & Literary Studies
Arab World English Journal
eISSN: 2550-1542

AWEJ for translation & Literary Studies Volume, 1 Number 4, October 2017 Pp. 151- 159
DOI: http://dx.doi.org/10.24093/awejtls/vol1no4.12

Machine Translation Quality of Khalil Gibran's The Prophet

Zakaryia Mustafa Almahasees
Department of European Languages ,School of Humanities
University of Western Australia, Perth, Australia

Abstract
Machine translation (MT) systems are widely used throughout the world freely or at low cost. The spread of MT entails a thorough analysis of translation produced by such translation systems. The present study evaluates the capacity of two MT systems-Google Translate and Microsoft Bing translator- in translation from Arabic into English of Khalil Gibran's literary masterpiece - The Prophet (2000). The question that arises in the study is could we trust MT in the translation of literary masterpieces across languages and particularly from Arabic to English? How close does MT output to human translation? To conduct that, the study is adopted Bilingual Evaluation Understudy (BLEU) of Papineni (2000). MT output analysis showed that MT is not accurate, intelligible and natural in translating literary texts due to the difficulty of literary texts, as they are full of metaphors and cultural specifications. Besides, there are some linguistic errors: lexical, syntactic and misinformation. The study also found that both systems provided similar translation for the same input due to either the use of similar MT approach or learning from previous translated texts. Moreover, both systems in some instances, achieve good results at the word level, but bad results at collocation units. The study also showed that automatic translation is insufficient for providing a full analysis of MT output because all automatic metrics are misleading due to dependence on text similarity to a reference human translation. For future research, the study recommended conducting a correlative study that combines manual and automatic evaluation methods to ensure best analysis of MT output. Machine Translation (MT) is still far from reaching fully automatic translation of a quality obtained by human translators.

Keywords: English-Arabic translation, Google Translate errors, Microsoft Bing errors, machine translation errors

Cite as: Almahasees, Z. M. (2017). Machine Translation Quality of Khalil Gibran's The Prophet. *Arab World English Journal for Translation & Literary Studies, 1*(4).
DOI: http://dx.doi.org/10.24093/awejtls/vol1no4.12

Introduction

Machine Translation (MT) is the subfield of computational linguistics that concern in the using of computer software in translation across human languages. Machine Translation is an interrelated field that covers three main interrelated knowledge disciplines: translation, linguistics and translations' software. In this regard, Almahasees (2017, p.1) presents a comprehensive definition of MT as "the use of computers in the process of translation from one natural language into another." He adds that the integration of Translation and technology has resulted in "automatic translation or translation by a computer called Machine Translation." He (2017, p.1) accentuates that such integration "allow computers to translate from one language into another."

Today, most of the scientific contribution is written in English, and thus globalization has brought English as an international language, which becomes the language of instruction in most global universities. In the case of people who do not master English, they tend to use MT as translation aid in their studies. Actually, MT is used widely, and there is a set of translation software that allowed the potential users to translate full documents, more quickly and freely or at low cost than human translators can. The most important aim of MT is to generate a translation that is similar to human translation, which is acceptable by human translators, clients, and readers. Trotsky (2016), product lead of Google Translate, shows that Google translate users extend more than 500 million per day. He mentions that most translation is conducted "Between English and Spanish, Arabic, Russian, Portuguese, and Indonesian." State page numbers with direct quotation

Machine Translation Evaluation (MTE) is considered as a primary step conducted by the designers of the systems to check the capacity, strength, potentiality, limitations, effectiveness and its attraction to the potential users. Historically, MT evaluation has proven difficult and continual struggle due to the lack of a method of evaluation that is accepted by all researcher. Essentially, the apparent results of evaluations make the potential users aware of MT feasibility.

Given above the spread of MT and the significance of MTE, the study answers the following questions, what is the degree of accuracy of MT in handling literary collocations? How close is MT translation to human translation? Which is the best system in rendering English literary collocations? What are the difficulties encountered by MT software in dealing with Arabic literary texts? The paper sheds lights on the usability and capacity of Google and Microsoft Bing in rendering Arabic literary collocations into English. Thus, the paper evaluates the MT output in comparison to human reference translation that is available for each chosen extracted collocations from Gibran's the prophet (2000).

Review of Literature

Several studies have been applied to measure and assess the quality and accuracy of machine translation by using automatic evaluation metrics such as WER, BLEU, METEOR, NIST, TER, ATEC, F-Function and others. Nießen et al. (2000) in An Evaluation Tool for Machine Translation developed an evaluative method to assess the intelligibility of MT systems. They introduced two tools for MT evaluation: WER (Word Error Rate) and SSER (Subjective Sentence Error Rate). They contended that WER and SSER tools as fast, semiautomatic, consistent, and suitable results. Furthermore, Melamed et al. (2003) in Precision and Recall of Machine Translation introduced Precision, Recall, and F-Measure. They believe that Precision, Recall, and F-Measure are more reliable than BLEU metric of Papinini et al. (2000) since their metrics are used for information retrieval, data mining and search engines. In a similar fashion, Moreover, Palmer (2005) in User-

Centered Evaluation for Machine Translation of Spoken Language developed the User-Centered method to assess MT, which is based on comparing MT output to human reference translation. Thus, he used his new method to test Arabic to English and Mandarin to English. He compares the output of MT in comparison to human reference translation and then asked experts to estimate and judge the proximity of MT to human reference translation.

In reference to Arabic Natural Processing, it started with a contribution of Akiba et al. (2006) in their Using Multiple Edit Distances to Automatically Grade Outputs from Machine Translation Systems designed an evaluation method that is relevant to Speech-to-Speech MT Systems (SSMT). Akiba's method is "Grader based on Edit Distance" that measure the score of MT output by using a decision tree. Therefore, they have conducted several tests, and they contended that Radar Based Edit Distance (RBED) is more accurate than BLEU. As stated above that the study adopts BLEU as evaluation metrics, Yang et al. (2008) in their Extending BLEU Evaluation Method of Linguistics Weight has conducted a study on BLEU and found that BLEU achieves the best results, especially through N-Gram weights.

Furthermore, Condon et al. (2012) in their study Evaluation of 2-way Iraqi Arabic–English speech Translation Systems using Automated Metrics assessed the way Iraqi –Arabic English are handled by MT systems. They found that high frequencies of MT errors in pronouns, subject pronoun inflection, word ordering and a very low frequency of errors in polarity errors. Moreover, Adly and Al-Ansary (2009) 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. They found that automatic evaluation metrics have not taken into account the importance of cohesion and semantic features of Arabic style; hence, more work needs to be done in the direction of translation between English<> Arabic. Alqudsi et al. (2014) in their Arabic Machine Translation survey discussed the obstacles for Arabic Machine Translation. The research found that it is hard to have a suitable machine translation that meets human requirements and expectations. Toral and Way (2015) in their Translating Literary Text between Related Languages using Statistical Machine Translation (SMT) assessed the translation of literary texts between Spanish and Catalan languages. They built their own system that based on state-of-the-art domain- adaptation techniques. They evaluate MT output in comparison to human professional translation. They evaluated MT output using BLEU, and they found that their system achieves better results "the translations produced by our best system are equal to the ones produced by a professional human translator in almost 20% of cases of an additional 10% requiring edits." As pointed out by Toral and Way (2015) Machine-Assisted Translation of Literary Text: A Case Study.

MT advancements proceed with a progressive manner. Currently, Neural Machine Translation (NMT) advances the field of MT to in preference to its previous statistical methods. Google and Microsoft Bing have started using NMT since September 2017 with an aim to advance MT to be in a way similar to human professional translation. Thus, the study adopts BLEU to assess how Google and Microsoft Bing in dealing with the translation of the literary texts from Arabic into English. BLUE, is "the best known and best-adapted machine evaluation for machine translation" as stated in Euro Matrix (2007). 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 the same text." (Al-kabi, 2013, p. 1) Vilar et al. (2006) confirm

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 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 to its best correlation with human judgments. This is the reason why the present study adopts BLEU.

Methodology

The present study adopts BLEU method to assess Google and Microsoft Bing outputs to find the effective and the best translation system in handling literary texts. The corpus of the study has selected from Gibran's The Prophet (2000) chapters. 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 reference human translation that is available at the English version of the book. The source sentence is extracted from Gibran's The Prophet chapters of the reference human translation that is available in English. The researcher analysed the MT output at the level of sentences to find the ngram strings in order to measure the precision of each sentence in comparison to human reference translation. In this respect, BLEU is calculated as follow:

Brevity Penalty BP= $\sum \quad e\left(1-\frac{r}{c}\right) \; if \; c < r \; or \; if \; c \leq r$

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

The present section illustrates the procedures of study's analysis and the ways of measuring MT outputs. The study is conducted to verify the usability and effectiveness of MT output through text similarity evaluation metric, BLEU. To illustrate how BLEU works, the following sentence explains how to measure the proximity between the candidate (MT output) and human reference translation. Our illustrative example is:

Source Text:

المحبة تكللكم فهي أيضا تطلبكم, وكما تعمل على نموكم هكذا تعلمكم وتستأصل الفاسد منكم

Google **(**Candidate 1): For as love also afflict you, it also hardens you, and as you work on your growth, so will you learn and eradicate the corrupt form you.

Bing (candidate 2): For even as love is crowned with you, it's also your stiffness, and as you grow your way, you teach and wipe out the rotten of you.

Reference: For even as love crowns you so shall he crucify you. Even as he is for your growth so is he for your pruning. (Project Gutenberg Australia, 2006)

The corpus of the research is analysed by using most widely known text similarity metric, BLEU. BLEU is an evaluation metric based on counting the similarity in terms of unigram (one word),

bigram (two words), trigram (three words), and tetragram (four words). Almahasees (2017, p.3) mentioned that "The similarity in number of unique words is complemented by a test of a similarity in strings of 2, 3, 4 words." Therefore, Papineni et al. (2000, p.1) indicated that the core idea of BLEU is "the closer a machine translation is to a professional human translation, the better it is."

Analysis and Discussion

Firstly, the analysis of the above example is conducted through counting the number of words of the candidates of Google and Bing along with a number of reference sentence words as shown in

Table 1. *The precision (PN) Value example (1)*

MT System N-gram P	Google	Microsoft Bing
PN 1	0.290	0.206
PN 2	0.066	0.035

Then, we analysed MT Candidates to identify n-gram and precision values against the reference translation, as for the above example illustrated below in Tables (2) and (3):

Table 2. *The precision Value of Example (2)*

MT system Ngram	Google	Bing
Unigram	8/31	7/29
Bigram	2/30	1/28
Trigram	0/29	0/27
Tetragram	0/28	0/26

After that, we measure the results of Brevity penalty for dividing the number n-grams of the reference sentence based on Brevity Penalty metric of Papineni et al. (2002) as follows:

$$\text{BP}= e\left(1-\frac{26}{31}\right)=1.17$$

$$\text{BP}= e\left(1-\frac{26}{29}\right)=1.10$$

The calculation shows that the Brevity Penalty for both Google and Microsoft Bing is (1) because the entire candidates are longer than the reference. In our example of Google, C= 31, R=26, and thus 31<26 then BP= 1 and for Bing C=29, R= 26 and when 26<29 then BP= 1. Alkabi et al. (2013) indicated that BP "n-gram precisions penalise candidate sentences found shorter than their reference counterparts, also it penalises candidate sentences, which have over generated correct forms." It has been also noticed that MT system in the sample only is capable of rendering single words, while collocations are absent from their translation.

More importantly, the analysis of the data shows that both systems produce the same translation for seven sentences of the study, as the following: capable of rendering single words, while collocations are absent in their translation.

جميل أن تعطي من يسألك ماهو في حاجة الية أن تعطي من اليسألك وانت تعرف حاجته

Google: It is nice to give someone who asks you what he needs. But more than that to give who does not ask you and you know his need.

Bing: It is nice to give someone who asks you what he needs. But more than that to give who does not ask you and you know his need.

Reference: It is well to give when asked, but it is better to give unasked, through understanding.

Table 3. *BLEU Score for both systems*

MT System **N-gram P**	**Google**	**Microsoft Bing**
Unigram	**6/30**	**6/30**
Bigram	**3/29**	**3/29**
Trigram	**1/28**	**1/28**
Tetragram	**1/27**	**1/27**

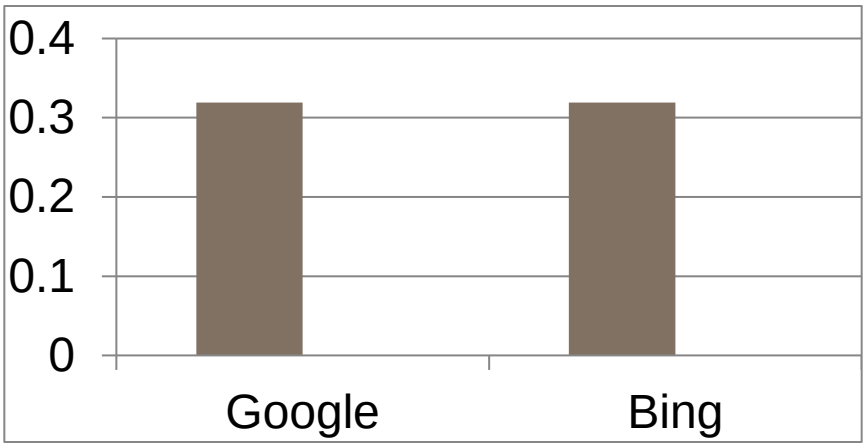

Figure 1. Precision Scores for both systems

In conclusion, the below chart elucidates the overall assessment of Google and Bing in translating literary texts from Arabic into English.

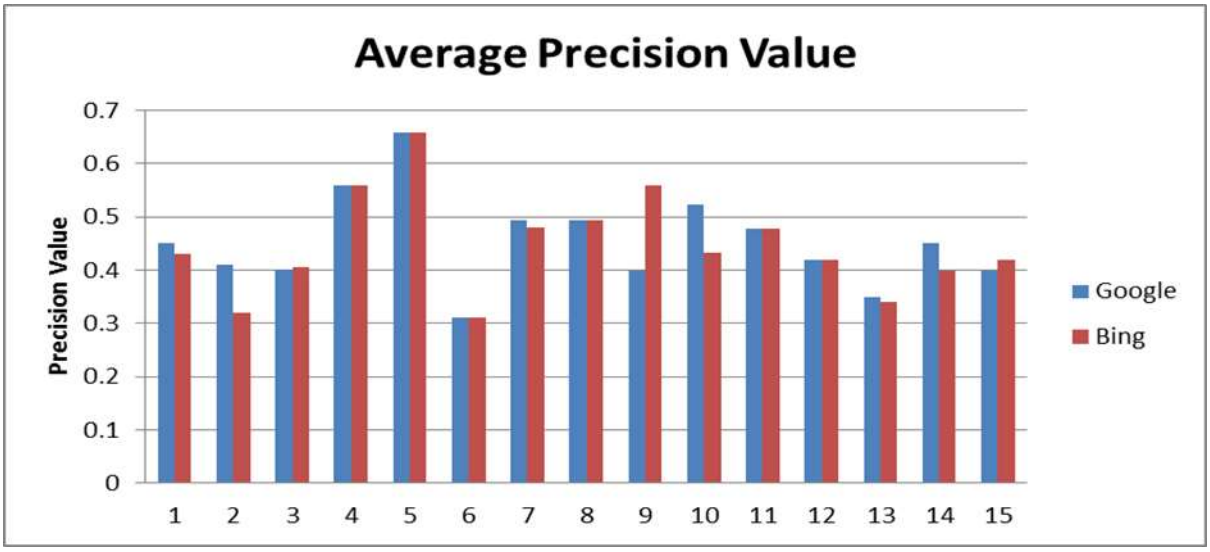

Figure 2. Average Precision Value

Conclusion

Translation from Arabic to English is a challenging task due to differences in syntactical, morphological and semantic features between the two languages. Both systems have problems in rendering literary texts due to the nature of literary language, which is full of connotations, culture specifications, literary personifications and images. The study also found that both systems provide similar translation for the same input, which is due to either the use of similar MT approach or learning from previous translated texts. Furthermore, both systems, in some cases, achieve good

results at the word level, but bad results at collocation combinations. Machine Translation is still far away from reaching translation quality that achieved by human translators.

Recommendations

The study recommends future research in the correlation between Automatic and manual evaluation methods, which would provide strong and more reliable results, regard the usability of MT systems. Moreover, it is recommended an analysis of MT output from a linguistics point of view on holistic and analytic levels to provide the Machine translation field with a rich literature, which helps 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.

About the author:

Zakaryia Mustafa Almahasees is a PhD student in Machine Translation at Department of European Languages, University of Western Australia, Perth, Australia. He is working on English-Arabic machine translation systems. He is a holder of an MA in English Language and Literature/ Literature from Jadara University, Jordan and B.A in English Language and Literature from Yarmouk University, Irbid, Jordan. His research interests focus on translation and technology, Machine Translation and Machine Translation Evaluation.

References:

Adly, N. & Ansary, S. (2009). Evaluation of Arabic Machine Translation System based on the Universal Networking Language. *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems* "NLDB 2009", Saarland University, Saarbrücken Germany, June 24-26.

Al-Kabi, M. N., Hailat, T. M., Al-Shawakfa, E. M., & Alsmadi, I. M. (2013). Evaluating English to Arabic machine translation using BLEU. *International Journal of Advanced Computer Science and Applications* (IJACSA), 4(1), 66-73.

Almahasees, Z. M. S. (2017). Assessing the translation of Google and Microsoft Bing in translating political texts from Arabic into English. *International Journal of Languages, Literature and Linguistics 3(1*), 1-4.

Alqudsi, A., Omar, N., & Shaker, K. (2014). Arabic machine translation: a survey. *Artificial Intelligence Review*, 1-24.

Condon, S., Arehart, M., Parvaz, D., Sanders, G., Doran, C., & Aberdeen, J. (2012). Evaluation of 2-way Iraqi Arabic–English speech translation systems using automated metrics. *Machine translation*, 26(1), 159-176.

Euro Matrix. (2007). Survey of Machine Translation Evaluation. Statistical and Hybrid Machine Translation between all European Languages. Retrieved from http://www.euromatrix.net/deliverables/Euromatrix_D1.3_Revised.pdf

Project Gutenberg Australia.(2006).*The prophet Translation*. Retrieved from http://gutenberg.net.au/ebooks02/0200061h.html

Khalil, Gibran. (2000). *The prophet*. Retrieved from http://gutenberg.net.au/ebooks02/0200061h.html

Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and recall of machine translation. *In Proceedings of 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--short papers- *V 2* (pp. 61-63). Association for Computational Linguistics.

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 39-45.

Palmer, D. D. (2005, March). User-centered evaluation for machine translation of spoken language. *In Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference* on V *(5),* pp. v-1013). IEEE.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the association for computational linguistics.* (ACL '02). Stroudsburg, PA, USA, pp. 311-18.

Toral, A., & Way, A. (2015). Machine-assisted translation of the literary text: A case study. *Translation Spaces*, 4*(2),* 240-267.

Trotsky, Barak. (2016). 11 Google Translate Facts You Should Know. Retrieved from http://www.k-international.com/blog/google-translate-facts /

Vilar, D., Xu, J., d'Haro, L. F., & Ney, H. (2006, May). Error analysis of statistical machine translation output. *In Proceedings of LREC* (pp. 697-702).

Y. Akiba , K. Imamura , E. Sumita , H. Nakaiwa , S. Yamamoto , H. G. Okuno. (2006) Using multiple edit distances to automatically grade outputs from Machine translation systems. *IEEE Transactions on Audio, Speech, and Language Processing*, v.14 *(2),* p.393-402, [doi>10.1109/TSA.2005.860770]

Yang, M., Zhu, J., Li, J., Wang, L., Qi, H., Li, S., & Daxin, L. (2008, November). Extending BLEU evaluation method with linguistic weight. *In Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference IEEE* (pp. 1683-1688).