An Automatic Evaluation for Online Machine Translation: Holy Quran Case Study

Emad AlSukhni
Computer Information Systems Department
Yarmouk University
Irbid, Jordan

Mohammed N. Al-Kabi
Computer Science Department
Zarqa University
P. O. Box 2000
13110 Zarqa -Jordan

Izzat M. Alsmadi
Computer Science Department
University of New Haven
West Haven
CT 06516, USA

***Abstract*—The number of Free Online Machine Translation (FOMT) users witnessed a spectacular growth since 1994. FOMT systems change the aspects of machine translation (MT) and the mass translated materials using a wide range of natural languages and machine translation systems. Hundreds of millions of people use these FOMT systems to translate the holy Quran (Al-Qurʾān) verses from the Arabic language to other natural languages, and vice versa. In this study, an automatic evaluation for the use of FOMT systems to translate Arabic Quranic text into English is conducted. The two well-known FOMT systems (Google and Bing Translators) are chosen to be evaluated in this study using a metric called Assessment of Text Essential Characteristics (ATEC). ATEC metric is one of the automatic evaluation metrics for machine translation systems. ATEC scores the correlation between the output of a machine translation system and professional human reference translation based on word choice, word orders and the similarity between MT output and the human reference translation. Extensive evaluation has been conducted on two well-known FOMT systems to translate Arabic Quranic text into English. This evaluation shows that Google translator performs better than Bing translator in translating Quranic text. It is noticed that the average ATEC score does not exceed 41% which indicates that FOMT systems are ineffective in translating Quranic texts accurately.**

***Keywords*—machine translation; language automatic evaluation; Statistical Machine Translation; Quran machine translation; Arabic MT**

I. Introduction

The Arabic word "Quran", "Qurʾan" or "(Al-Qurʾān)" means literally "the recitation". The holy Quran is considered by Muslims around the world as the verbatim word of God (Allah) dictated by Allah through the archangel Gabriel (Jibrīl) to the Prophet Muhammad. Holy Quran is divided into 114 Sûrats (chapters). Each Sûrat consists of several number āyāts (verses). The length of these Sûrats varies considerably, where Sûrat length is measured in the number of āyāts that it has and that varies from few āyāts (verses) as in the first Sûra (Al-Fatiha) which consists of seven verses only to hundreds of āyāts within the same Quranic Sûrats such as the second Sûra (Al-Baqarah) that consists of 286 verses. The total number of Quranic verses is 6,236 [1-4].

The Quran is originally saved and written in the Arabic language. To sustain its high authenticity, translators of Quran into other natural languages strive to be consistent in terms of such translation. Nonetheless, it is known that when translating words and statements from one language to another, more than one word can be used to translate/interpret a particular word. In addition, and due to the nature of differences between the different natural languages, in many cases, word-by-word translation may not produce meaningful statements. As such, the general statement meaning or interpretation may be necessary to explain the meaning of Quran statements despite the fact that such meaning translation or interpretation may not be literal or word by word identical to the original one.

The expansion of information through the Internet gives a huge source of information to all humans worldwide to access and read information. Information can be then written in one language in one place in the world, read and translated to all other languages and places. Many information retrieval systems and search engines such as Google provide their own interpretation of information and data posted to web pages when translating text from one language to another. The correctness or accuracy of such online dictionary or language translation should then be evaluated especially when it comes to sensitive texts such as holy or religious books including the holy Quran.

The Arabic language gains interest recently due to many religious and political factors. Non-Arabic speakers still need to translate the holy Quran in order to know its meaning or to understand it. Many none Arabic-speakers use online FOMT systems to translate the holy Quran. These systems need to be evaluated to know the quality of their outputs.

The evaluation of MT systems is subjective and not objective whether it is conducted by a professional human expert or by computerized systems using metrics such as BLEU or ATEC. The quality of automatic machine translation may vary due to different algorithms used by these MT systems. The variations in the evaluation of different automatic metrics to evaluate the effectiveness of different machine translation systems is due to the following reasons. First, it may depend on the machine translation algorithms, dictionary size or correctness, or the nature of the language and the input text. The evaluation of the quality of translation by ATEC (Assessment of Text Essential Characteristics: http://mega.ctl.cityu.edu.hk/ctbwong/ATEC/ [5]) metric is based on Word choice and word position to automatically evaluate MT systems. ATEC metric is proposed by Wong and Kit and was presented for the first time in 2008 [6]. ATEC is used in this study to evaluate the effectiveness of two Well-known FOMT online systems (Google and Bing Translators).

The rest of the paper is organized as follows: Section II exhibits an overview of the related work. Section III presents the proposed methodology. Section IV exhibits the experiments, and section V presents the results of the conducted experiments on four Sûrats (Chapters) from the holy Quran which totally constitute to around 9% of the whole number of Quranic verses. Section VI presents the conclusions.

II. Related Work

The cost and speed are the main reasons that lead to the widespread and extensive use of machine translation systems. No one claims that the MT systems and their automatic evaluation are better than human professionals. However, these tools are cheaper and faster when compared with human effort and translation. The literature has a large number of papers that discussed different techniques to automatically evaluate MT systems. Some of these studies will be presented in this section.

Bi-Lingual Evaluation Understudy (BLEU) metric is one of the earliest and widespread automatic evaluation metrics which is presented in 2002 by [7]. BLEU popularity is due to the easiness of its computation. In addition, it can be considered as language independent. Many versions of this metric are presented in different studies. BLEU uses a modified unigram precision to compute the correlation between candidate and reference translations.

(METEOR: Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005 [8]; Lavie and Denkowski, 2009) is a recall-oriented metric that tries to solve some of the problems in BLUE such as dealing with an exact matching of words' locations. The formula can deal with more than one reference statement for evaluation. Similar to ATEC approach, METEOR performs words' alignments when the word order in reference translations is different from word order in candidate translations.

Wong and Kit presented in their paper [6] for the first time their novel automatic metric which is called ATEC. In order to express an idea in any natural language, you have to choose the right words and put them in the right order. This simple fact of natural language research is used by Assessment of Text Essential Characteristics (ATEC). Authors Wong and Kit [5, 6] presented ATEC (F-measure oriented metric) for machine translation. As discussed earlier, ATEC metric computes the precision and recall in order to compute F-measure. Afterward, the penalty is computed which is based on the differences in word orders. The final value of ATEC is calculated by multiplying F-measure by the penalty as shown in the introduction section of this study. This is the version which is adopted in this study. The same authors of ATEC [5, 6] proposed later an enhanced ATEC version of this metric and presented an updated version of ATEC in [10]. This new metric includes word position and information flow.

The study of Al-Kabi, Hailat, Al-Shawakfa, and Alsmadi [11] uses BLEU metric to evaluate the quality of two free online translators (Google and Babylon translators) in translating English sentences to Arabic. They conclude in their study that the quality of Google translator is better than the quality of Babylon translator. Al-Deek, Al-Sukhni, Al-Kabi, and Haidar [12] conducted a study to automatically compare the translation quality of Google Translator and IM Translator) by using ATEC metric. Their results showed that Google translator outperforms IM translator in terms of the quality of translating English sentences to Arabic.

The study of Hadla, Hailat, and Al-Kabi [13] also evaluates the translation quality of two free online machine translators (Google and Babylon translators) from Arabic to English. In their study, the researchers select BLEU metric to measure the quality of translation of the two systems under evaluation. The researchers used a corpus that consists of more than 1000 Arabic sentences in their study. The result of their study also indicates that the quality of Google translator is better than the translation quality of Babylon. The corpus of Arabic sentences they used with the English translations of these sentences consists of 4169 Arabic words, where the number of unique Arabic words is 2539. This corpus is released online to be used by other researchers. These Arabic sentences were distributed among four basic sentence functions (declarative, interrogative, exclamatory, and imperative). Hadla, Hailat, and Al-Kabi extended their study in [14]. They used METEOR 1.5 and BLEU metrics to evaluate the effectiveness of Google Translate and Babylon MT systems, using 1033 Arabic sentences with two English reference translations for each Arabic sentence. A system is built to automatically evaluate MT using METEOR 1.5 and BLEU. As their previous study the Arabic sentences are distributed equally among four basic sentence functions (declarative, interrogative, exclamatory, and imperative). The results of study [14] showed clearly that the outputs of Google Translate are closer to reference translations than Babylon MT system, except in the case of translating exclamatory Arabic sentences to English where Babylon MT system outputs prove to be closer to reference translations than the outputs of Google Translate.

III. Methodology

This section presents the methodology followed in this paper to evaluate two FOMT systems. As shown in Fig. 1, the study starts with the collection of the dataset that contains the original Arabic texts of four Quranic chapters of the holy Quran, where the total number of verses of these four chapters is 486 as shown in Table I, in addition to the English translation of each collected Arabic verse of the four selected chapters. This study selects the four chapters with different lengths (i.e. number of verses in the chapter) in the holy Quran. Table I shows the number of verses in each chapter used in this study. The translated version of these Quranic chapters was collected from King Fahd Complex for the Printing of the Glorious Quran website [15].

TABLE I. The Sizes of Used Quranic Chapters

Chapter Name (AR)	**Chapter Name (EN)**	**Number of Verses**
Al-Fâtihah	The Opening	7
Al-Baqarah	The Cow	286
Al-Kahf	The Cave	110
Yâ-Sîn	YaSin	83
Total		486

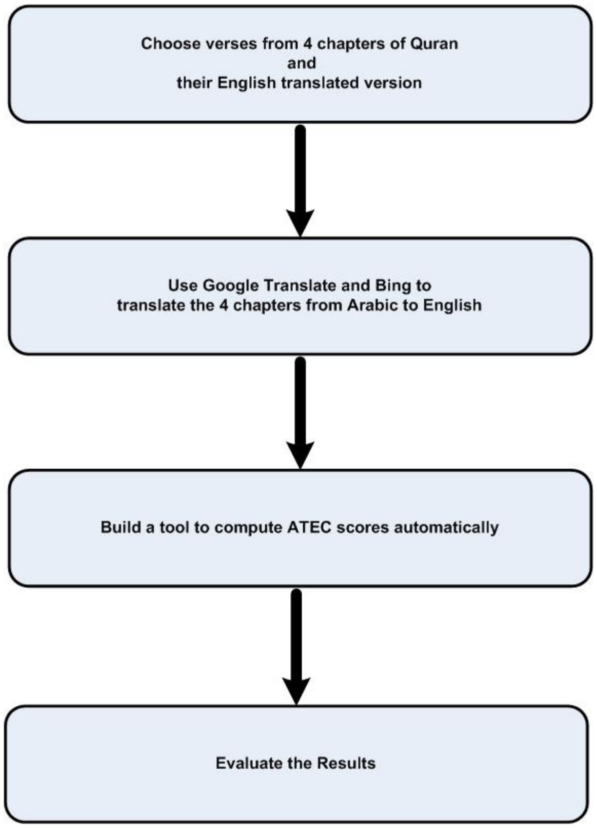

Choose verses from 4 chapters of Quran
and
their English translated version
Use Google Translate and Bing to
translate the 4 chapters from Arabic to English
Build a tool to compute ATEC scores automatically
Evaluate the Results

Fig. 1. Schematic overview of the methodology

As shown in Fig. 1, in the second phase of the methodology of this study, the collected Quranic verses are translated from Arabic to English using both FOMT systems under evaluation (Google and Bing Machine Translators). After translating the collected verses, ATEC scores should be computed based on the correlation between the automatically translated verses (i.e. output of FOMT systems) and those translated by professional humans (i.e. verses collected from human translated version of the Glorious Quran).

ATEC score computation is based on unigram F-measure to measure the word choice (i.e. matching between candidate and reference translations). Therefore, this metric computes Precision (P) and Recall (R) in order to compute afterward the unigram F-measure which measures word choice as shown in the following formulas 1 and 2 [6]:

$$P(c,r)=\frac{M(c,r)}{|c|} \tag{1}$$

$$R(c,r)=\frac{M(c,r)}{|r|} \tag{2}$$

Where P(c, r), R(c, r), and M(c, r) are: precision, recall and number of matched unigrams between the candidate (c) and reference translations respectively. |c| and |r| are the lengths of candidate translation and reference translation respectively.

Next unigram F-measure (F(c, r)) can be computed as shown in formula 3 [6]:

$$F(c,r)=\frac{2\times P(c,r)\times R(c,r)}{P(c,r)+R(c,r)} \tag{3}$$

As mentioned before, this metric is based on word choice which measured as described in formula 3. Afterward, this study measures the word order in terms of penalty rate (Penalty) which is based only on word position differences between a candidate translation and one or more reference translations as shown in formula 4 [6]:

$$Penalty=\begin{cases}1-(PosDiff\times 4) & if\ \ PosDiff\le 0.25\\ \\ 0 & if\ \ PosDiff>0.25\end{cases} \tag{4}$$

Afterward, the ATEC measure is computed using F-measure (formula 3) and the penalty (formula 4). Formula 5 presents the ATEC metric [6]:

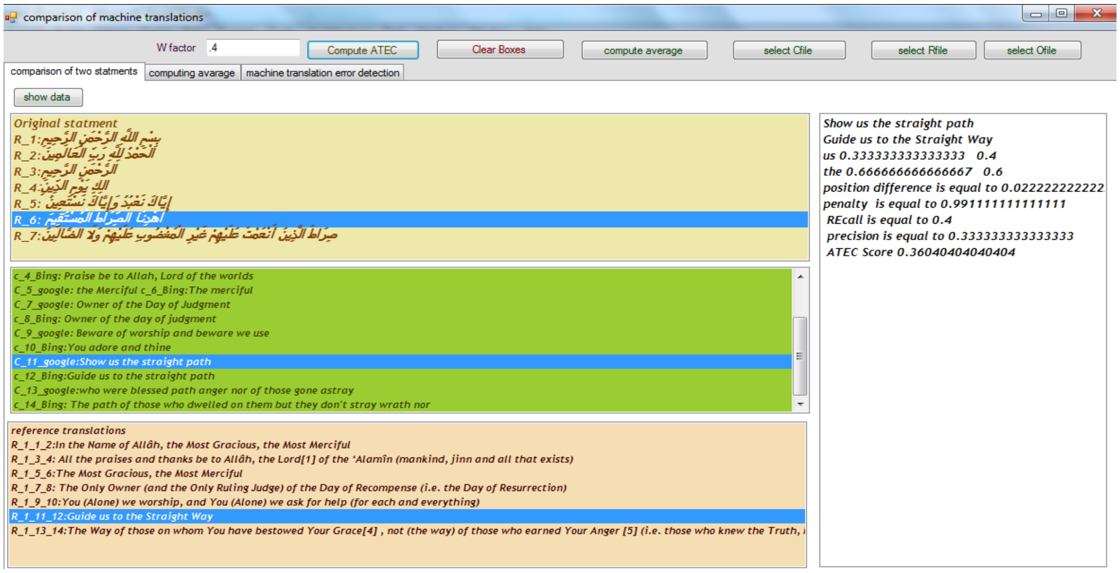
$$ATEC=F(c,r)\times Penalty \tag{5}$$

ATEC score is calculated sentence by sentence and then averaged for the whole document.

To achieve the objectives of this research, a tool is built to automatically compute ATEC score for each FOMTs under evaluation (Google and Bing Machine Translators). In the last phase of this study, the results are evaluated (i.e. ATEC score for each FOMT system). The results are used as a sign to tell us about the quality of Arabic to English translation produced by each FOMT system. Moreover, the collected results are used to indicate which translator out of FOMTs under evaluation (Google and Bing Machine Translators) is better. thus, an average ATEC score is produced by computing the average of ATEC scores for all verses in each chapter for each FOMT system.

IV. Experiments

To automatically evaluate the performance of machine translation, an evaluation tool using C# is developed to compute ATEC score(s). The main screen of the tool is shown in Fig. 2. This tool produces the evaluation of the translation either for one verse (as shown in the Fig. 2) or for the complete chapter. Fig. 2 shows the evaluation of the sixth verses in Sûrat Al-Fâtihah. Fig. 2 shows clearly the values of Precision (P) and Recall (R), and ATEC score.



Original statment
R_1:بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ
R_2:الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ
R_3:الرَّحْمَنِ الرَّحِيمِ
R_4:الِكِ يَوْمِ الدِّينِ
R_5: إِيَّاكَ نَعْبُدُ وَإِيَّاكَ نَسْتَعِينُ
R_6: اهْدِنَا الصِّرَاطَ الْمُسْتَقِيمَ
R_7:صِرَاطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ الْمَغْضُوبِ عَلَيْهِمْ وَلَا الضَّالِّينَ

c_4_Bing: Praise be to Allah, Lord of the worlds
C_5_google: the Merciful c_6_Bing:The merciful
C_7_google: Owner of the Day of Judgment
c_8_Bing: Owner of the day of judgment
C_9_google: Beware of worship and beware we use
c_10_Bing:You adore and thine
C_11_google:Show us the straight path
c_12_Bing:Guide us to the straight path
C_13_google:who were blessed path anger nor of those gone astray
c_14_Bing: The path of those who dwelled on them but they don't stray wrath nor

reference translations
R_1_1_2:In the Name of Allâh, the Most Gracious, the Most Merciful
R_1_3_4: All the praises and thanks be to Allâh, the Lord[1] of the 'Alamîn (mankind, jinn and all that exists)
R_1_5_6:The Most Gracious, the Most Merciful
R_1_7_8: The Only Owner (and the Only Ruling Judge) of the Day of Recompense (i.e. the Day of Resurrection)
R_1_9_10:You (Alone) we worship, and You (Alone) we ask for help (for each and everything)
R_1_11_12:Guide us to the Straight Way
R_1_13_14:The Way of those on whom You have bestowed Your Grace[4] , not (the way) of those who earned Your Anger [5] (i.e. those who knew the Truth,

```
Show us the straight path
Guide us to the Straight Way
us 0.333333333333333   0.4
the 0.666666666666667   0.6
position difference is equal to 0.022222222222
penalty  is equal to 0.991111111111111
 REcall is equal to 0.4
 precision is equal to 0.333333333333333
 ATEC Score 0.36040404040404
```

Fig. 2. The main screen of the Arabic ATEC-based FOMT Evaluation System

V. RESULTS

To extensively evaluate the quality of FOMT systems in translating Quranic text, the experiments were conducted on four chapters. Table 2 shows initial results of these experiments. Table 2 shows a number of verses in each chapter and the average ATEC score of each FOMT system for that chapter. Average ATEC score is produced by computing the average of ATEC score of all verses in each chapter for each FOMT system). The average ATEC score represents the quality of translation on a scale from 0 (the lowest quality) to 1 (the highest quality). Table II and Fig. 3 show the average ATEC score of Google translator varies from 0.34 (in The Opening (Sûrat Al-Fâtihah) chapter) to 0.41 (in The Cow (Sûrat Al-Baqarah) chapter) and the average ATEC score of Bing translator varies from 0.23 (in The Opening (Sûrat Al-Fâtihah) chapter) to 0.34 (in The Cow (Sûrat Al-Baqarah) chapter). Thus, it is concluded that Google Translate system is better than Bing translator in translating Quranic text.

TABLE II. AVERAGE ATEC SCORE OF EACH FOMT SYSTEM

Chapter	Average ATEC Score	
	Google MT	**Bing MT**
The Opening, (Al-Fâtihah)	0.34	0.23
Al-Baqarah	0.41	0.34
Al-Kahf	0.34	0.29
Yâ-Sîn	0.38	0.30

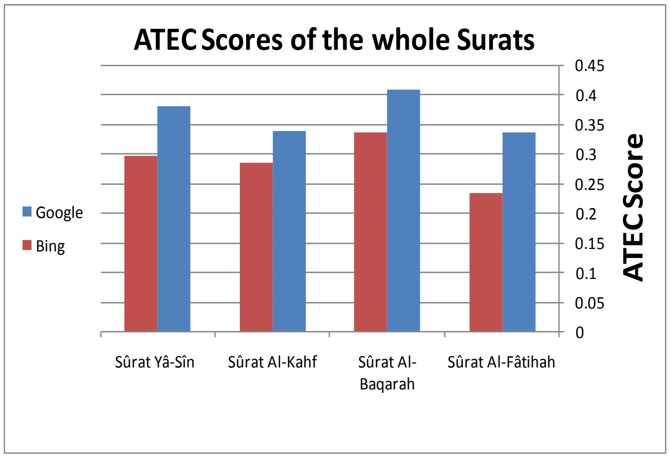


Fig. 3. Average ATEC score for each translated Sûrat on each FOMT

The results show that, although Google Translate system is better than Bing, yet the average ATEC score for both translators is less than 41% which is generally low.

In order to extensively investigate the quality of translation for both online systems, the developed tool was designed to give us the average ATEC scores for each FOMT system used to translate verses with different lengths (i.e. number of words in the translated verse). Fig. 4, Fig. 5 and Fig. 6 showed the average ATEC score for the translated verses based on their length in Sûrat Al-Baqarah, Sûrat Al-khahf and Sûrat YaSin chapters respectively.

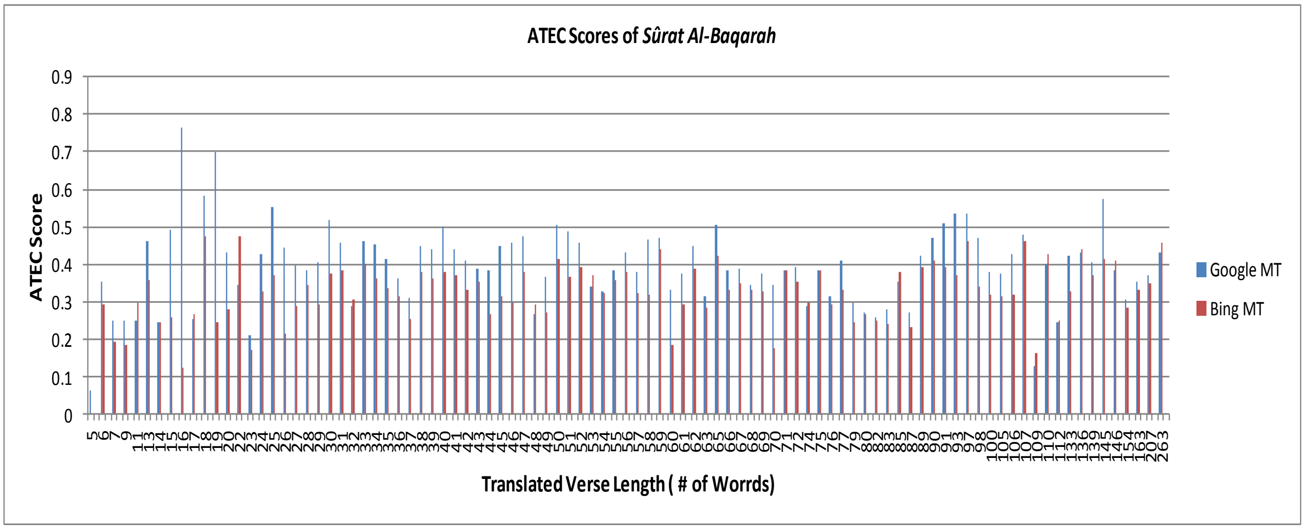


Fig. 4. Average ATEC score for different translated verses of The Cow (Al-Baqarah) chapter by the two FOMT Systems (classified by verse length)

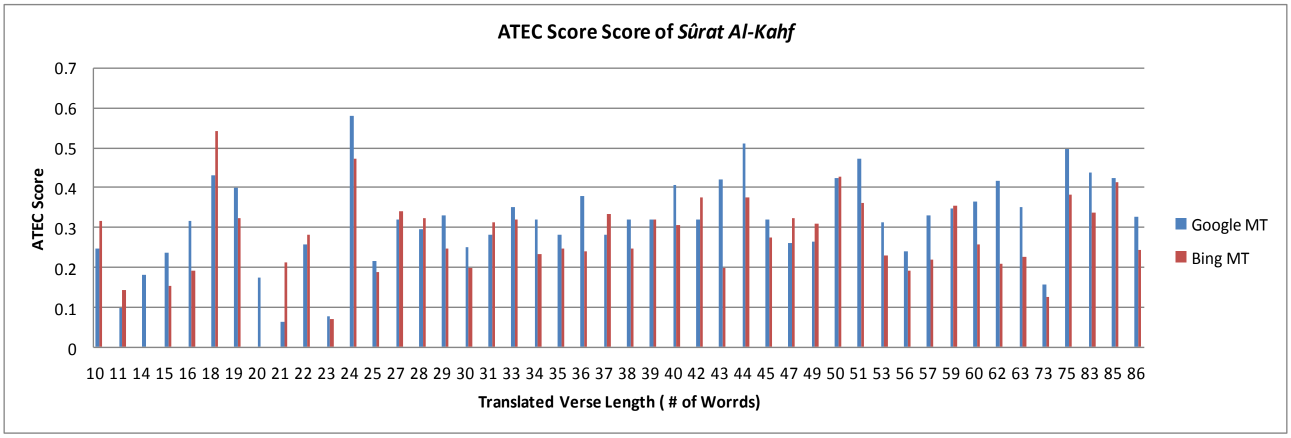


Fig. 5. Average ATEC score for deferent translated verses of Al-khahf on each FOMT (classified by verse length)

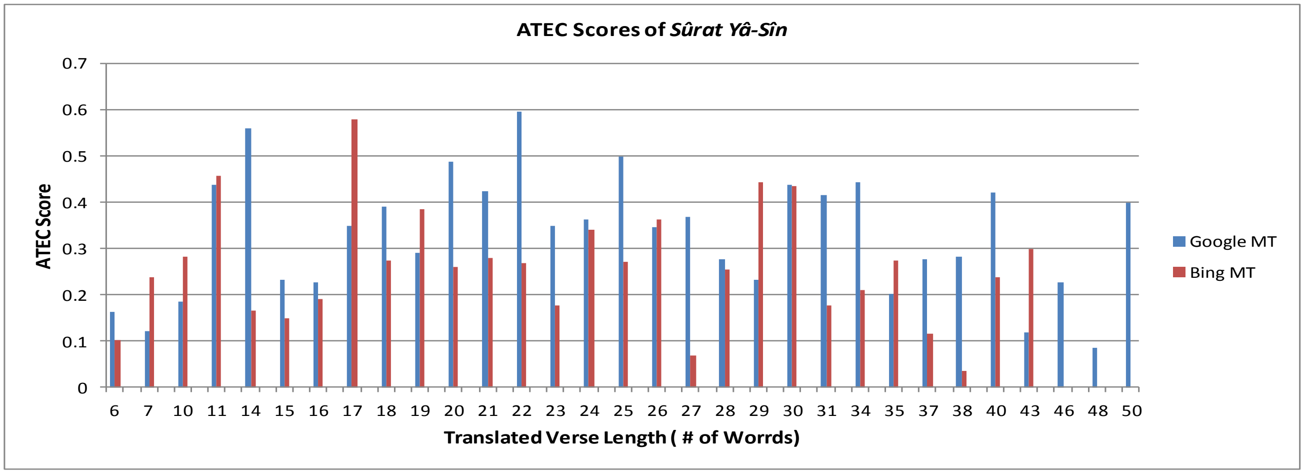


Fig. 6. Average ATEC score for deferent translated verses of Sûrat Yâ-Sîn on each FOMT (classified by verse length)

In most cases the translation of Google was better than the translation of Bing according to ATEC scores as shown in Fig. 4, Fig. 5 and Fig. 6. Moreover, from those Figures, it can be seen that the quality of the translation (which is presented by ATEC Scores) is independent of the length of the verse. The figures show that the ATEC scores of some translated verses with the short lengths are higher than the ATEC scores of some translated verses with the longer lengths and vice versa. From this observation, it is believed that the quality of verse translation is dependent on the words that composed the verse.

During the evaluation, it is found out that both FOMT systems have completely failed to translate many words (such as الأذقان(chins) and فعززنا (reinforced) in YaSin Chapter (Sûrat Yâ-Sîn), and like عسرا (distress) in (Sûrat Al-Kahf). It has also been noticed that in such cases, Google Translate System kept the word as is (in Arabic) in its translation while Bing translator used English characters to transliterate instead of translating it.

VI. Conclusion and Future Work

The need to translate Arabic text to other natural languages especially English is continuously growing. In this study, an automatic evaluation for two FOMT systems to translate Arabic Quranic text to English is conducted. The two Well-known FOMT systems: Google and Bing Translators are chosen to be evaluated in this study using ATEC. ATEC metric is one of the automatic evaluation metrics for machine translation systems. ATEC computes the correlation between the output of machine translation system and professional human reference translation based on word choice, word order, and the similarity between MT output and the human reference translation. A tool to compute ATEC score is built and used to evaluate the translation quality of the FOMT systems with a case study of four chapters of the holy Quran. The results showed that outputs of Google Translate System are better than the outputs of Bing translator. On the other hand, the average ATEC score for each translator did not exceed 41% which is generally very low especially for translating holy texts. The authors think that using look-up tables by MT systems to translate holy texts yield outputs that are fully matched with the best human translators.

As the issue of Quran translation is important to many readers and as there can be a significant difference in the meaning between “literal” and “contextual” translations, we think that a special framework and methods should be used to translate holy books in general. Our plan in future is to integrate results from automatic translations with domain experts (i.e. religious scholars). Many religious scholars for example prefer to use the word “interpretation” rather than “translation” to reflect the fact that the translation of Quran from Arabic to English or any other language can never render the exact original meanings and contexts.

References

[1] Quran. Retrieved April 01, 2013, From http://en.wikipedia.org/wiki/Quran

[2] What is Quran? Retrieved April 01, 2013, From http://www.cometoislam.com/quran.htm

[3] What is Quran? Allah describes the Quran in the Quran. Retrieved April 01, 2013, From http://www.iqrasense.com/quran/what-is-quran-allah-describes-the-quran-in-the-quran.html

[4] Islam. Retrieved April 01, 2013, From http://en.wikipedia.org/wiki/Islam

[5] B. Wong, and C. Kit, "ATEC: automatic evaluation of machine translation via word choice and word order," Machine Translation, vol. 23 No. 2-3, pp. 141–155, September 2009.

[6] B. T-M Wong, and C. Kit, "Word choice and Word Position for Automatic MT Evaluation," In Proceedings of the AMTA 2008 Workshop: MetricsMATR, 3 pages, Waikiki, Hawai'i, October, 2008.

[7] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02), Philadelphia, pp. 311-318, July 2002.

[8] S. Banerjee, and A. Lavie, "Meteor: An automatic metric for MT evaluation with improved correlation with human judgments," In Proceedings of ACL Workshop on Intrinsic & Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65-72, 2005.

[9] A. Lavie, and M. J. Denkowski, "The METEOR metric for automatic evaluation of machine translation," Machine Translation, 2009, vol..23, pp. 105–115, 2009.

[10] B. T-M Wong, and C. Kit, "The parameter-optimized ATEC metric for MT evaluation," In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 360-364, 2010.

[11] M. N. Al-Kabi, T. M. Hailat, E. M. Al-Shawakfa, and I. M. Alsmadi, “Evaluating English to Arabic Machine Translation Using BLEU,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 1, pp. 66-73, 2013.

[12] H. Al-Deek, E. Al-Sukhni, M. Al-Kabi, M. Haidar, “Automatic Evaluation for Google Translate and IMTranslator Translators: An Empirical English-Arabic Translation,” The 4th International Conference on Information and Communication Systems (ICICS 2013). ACM, Irbid, Jordan, 2013.

[13] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Evaluating Arabic to English Machine Translation," International Journal of Advanced Computer Science and Applications, Vol. 5, No. 11, pp. 68-73, 2014.

[14] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Comparative Study between METEOR and BLEU Methods of MT: Arabic into English Translation as a Case Study," International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher, Vol. 6, No. 11, pp. 215-223, 2015.

[15] http://qurancomplex.gov.sa/Quran/Targama/Targama.asp?TabID=4&SubItemID=1&l=eng&t=eng&SecOrder=4&SubSecOrder=1 accessed in 15/10/2013.

Authors Profile

Emad Mahmoud Alsukhni obtained his PhD in from Ottawa University in Canada in (2011), he obtained his Masters' degree in Computer and Information Science from Yarmouk University in (2003), and obtained his Bachelor degree in Computer Science from Yarmouk University in (2003). Alsukhni is an assistant professor at the Faculty of Information Technology and Computer Science at Yarmouk University in Jordan. Alsukhni research interests include Computer Networks, Information Retrieval, Sentiment analysis and Opinion Mining, and Data Mining.. He is the author of several publications on these topics.

Mohammed Al-Kabi, born in Baghdad/Iraq in 1959. He obtained his Ph.D. degree in Mathematics from the University of Lodz/Poland (2001), his master's degree in Computer Science from the University of Baghdad/Iraq (1989), and his bachelor degree in statistics from the University of Baghdad/Iraq(1981). Mohammed Naji Al-Kabi is an assistant Professor in the Faculty of IT, at Zarqa University. Prior to joining Zarqa University, he worked many years at Yarmouk University in Jordan, Nahrain University and Mustanserya University in Iraq. He also worked as a lecturer in PSUT and Jordan University of Science and Technology (JUST). Al-Kabi's research interests include Information Retrieval, Sentiment analysis and Opinion Mining, Big Data, Web search engines, Machine Translation, Data Mining, & Natural Language Processing. He is the author of several publications on these topics. His teaching interests focus on information retrieval, Big Data, Web programming, data mining, DBMS (ORACLE & MS Access).

Izzat Alsmadi is an associate professor in the department of computer science at University of New Haven. He obtained his Ph.D degree in software engineering from NDSU (USA), his second master in software engineering from NDSU (USA) and his first master in CIS from University of Phoenix (USA). He had a B.sc degree in telecommunication engineering from Mutah university in Jordan. He has several published books, journals and conference articles largely in software engineering, data mining, IR and NLP.