Analysing Patterns of Errors in Neural and Statistical Machine Translation of Arabic and English

ABSTRACT

This paper provides a comparative analysis of two Machine Translation (MT) engines: Google Translate (GT) and Microsoft Bing (MB). Previously, these MT engines adopted the statistical approach in their system. However, they are currently using the latest neural approach in their engines, which has become a trend in the MT field. The present data discusses the quality of the outputs by comparing the previous data using the SMT engines with the current data from the NMT engines. This paper also analyses the patterns of errors that exist in the MT outputs, which were generated using four Arabic texts and 5 English texts. Results reported a significant decrease of 72.2% and 73.1% in the number of errors found in GT and MB, respectively, with most of them are syntactic errors and incorrect terms. Missing conjunctions and determiners were also reported to be common mistakes in the analysis. Generally, the adequacy of both NMT engines has improved for English-Arabic language pairs. Even though the errors still exist, most of them can be easily corrected if thoroughly revised.

Keywords: Arabic, English, error analysis, machine translation quality, neural machine translation, statistical machine translation.

PROBLEM STATEMENT

Machine Translation (MT) has improved immensely since it was first introduced in the Georgetown-IBM experiment in the 1950s. However, adequacy errors remain problematic and require further studies and development. Moreover, many factors can contribute to the existing errors, such as language pairs, sentence length, text genres, types of MT systems and users[1]. Translation quality, on the other hand, is considered as "a subjective process which relies on human judgments"[2]. Therefore, many studies[3] opted for different approaches to measure the MT quality, such as edit distance, automatic MT evaluation metrics and error analysis.

Many developers nowadays have adapted the neural approach (Neural Machine Translation or NMT) in their MT engines. Compared to the previous statistical approach (Statistical Machine Translation or SMT), the current NMT engines have drawn increasing interest in both academic and professional communities, as many studies[4] have shown positive results using automatic MT evaluation. However, the abundance of errors in some results requires in-depth error analysis on the current NMT engines to understand the patterns of MT errors, which can be used as a guideline when post-editing the MT outputs. Hence, the focus of the present study is analysing patterns of errors in two NMT engines: Google Translate (GT) and Microsoft Bing (MB) and comparing the current results to the previous data collected by Haji Sismat[5].

---

[1] Guerberof, A.A. (2012). *Productivity and quality in the post-editing of outputs from translation memories and machine translation* (Doctoral dissertation). Universitat Rovira I Virgili, Tarragona, Spain; Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In *Proceedings of WPTP,* 11-20; Koponen, M., & Salmi, L. (2015). On the correctness of machine translation: A machine translation post-editing task. *The Journal of Specialised Translation,* 23, 118-136.

[2] Secară, A. (2005). Translation evaluation: A state of the art survey. In *Proceedings of the eCoLoRe/MeLLANGE workshop*, 39.

[3] Papipeni, 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 on association for computational linguistics,* 311-318; Specia, L, & Farzindar, A. (2010). Estimating machine translation post-editing effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10).* 33-41; Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In *Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice,* 2, 63-71.

[4] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations 2015,* San Diego, California, 1-15; Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., ... Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers,* 131-198; Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems,* 1-11.

[5] Haji Sismat, M. A. (2016). *Quality and productivity: A comparative analysis of human translation and post-editing with Malay learners of Arabic and English* (Doctoral dissertation). University of Leeds, Leeds, United Kingdom; Haji Sismat, M. A. (2019a). Neural and Statistical Machine Translation: A comparative error analysis. In *Proceedings of 17th International Conference of Translation,* 393-403; Haji Sismat, M. A. (2019b). Inverse Translation Quality: A comparative analysis between human translation and post-editing. *Journal of Arabic Linguistics and Literature*, 2, 91-105.

Another aspect of this study is the language pair: Arabic and English. These languages are distinctively different in many ways, such as the alphabet, grammar, and lexicons. For example, the grammatical numbers in English are typically described as singular and plural. However, Arabic also has dual nouns, which are usually followed by dual verbs in nominal sentences, such as الوالدان يعلمان أبناءهما (The parents are teaching their kids). In this example, the distinction between the two languages indicates no dual nouns, pronouns or verbs in the English sentences when compared to the Arabic sentence. Therefore, users, researchers and developers must be aware of these differences because MT engines may not be able to differentiate between them. Therefore, this paper also aims to provide an insight for developers on how to improve their MT engines, as well as evaluate the quality of the NMT outputs when compared to the previous data. Therefore, the current data is analysed to show the improvements made on current NMT engines without overselling the potential of the MT products, as suggested by Castilho et al[6].

RESEARCH QUESTIONS

The focus of this research is to analyse the patterns of errors in the NMT systems and then compare them to the ones in the SMT systems. Therefore, the present study attempts to answer the following research questions:

1. Has the quality of the two MT engines (GT and MB) increased using the neural-approach when compared to the previous data collected from SMT?
2. What are the patterns of errors currently exist in both GT and MB? Have the patterns of errors changed in both GT and MB?

METHODS

To find out whether or not the MT quality has increased, the present study uses the same nine technical texts (journalistic and legal) used in the previous studies[7]: five English texts and four Arabic texts, ranging from 151-311 words. Like the previous research, the present study selected two MT engines: Google Translate (GT) and Microsoft's Bing Translator (MB). The present study also focuses on the Arabic-English language pair as one of the aspects explored in the analysis.

[6] Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics,* 108(1), 109-120.

[7] Haji Sismat, M. A. (2016). *Quality and productivity: A comparative analysis of human translation and post-editing with Malay learners of Arabic and English* (Doctoral dissertation); Haji Sismat, M. A. (2019a). Neural and Statistical Machine Translation: A comparative error analysis*,* 393-403; Haji Sismat, M. A. (2019b). Inverse Translation Quality: A comparative analysis between human translation and post-editing, 2, 91-105.

In the error annotation phase, the present study adopted MeLLANGE error typology. According to MeLLANGE[8], the error typology has two categories: content- and language-related errors, followed by its subcategories. The present study does not intend to provide an exhaustive list of types of errors but to classify and quantify the number of errors so that it can investigate the error frequencies contained in the MT outputs. Even though many studies[9] adopted different error typologies for their analyses, the lists of types of errors are similar as these errors are only categorised or subcategorised differently. The reason for choosing MeLLANGE error typology is its comprehensive list of types of errors that can be applied to assessing both English and Arabic translations as previously carried out by Haji Sismat[10].

Table 1

MeLLANGE Error Typology

Content-transfer	Language
❖ Omission ❖ Addition ❖ Distortion in meaning ❖ SL intrusion: ➢ Untranslated translatable ➢ Too literal ➢ Units of weight/measurement, dates and numbers	❖ Syntax ❖ Wrong preposition ❖ Inflection and agreement: ➢ Tense/aspect ➢ Gender ➢ Number ❖ Terminology and lexis: ➢ Incorrect ➢ Term translated by non-term ➢ Inconsistent with glossary ➢ Inconsistent within TT ➢ Inappropriate collocation ❖ Hygiene: ➢ Spelling

[8] MeLLANGE. (2007). MeLLANGE: Multilingual eLearning in LANGuage Engineering.

[9] Izwaini, S. (2006). Problems of Arabic machine translation: evaluation of three systems. *The British Computer Society (BSC),* London, 118-148; Al-Samawi, A. M. (2014). Language errors in machine translation of encyclopaedic texts from English into Arabic: the case of Google Translate. *Arab World English Journal*, 182-211; Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., ... & Oflazer, K. (2014). Large scale arabic error annotation: Guidelines and framework. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference,* 2362-2369.

[10] Haji Sismat, M. A. (2016). *Quality and productivity: A comparative analysis of human translation and post-editing with Malay learners of Arabic and English* (Doctoral dissertation); Haji Sismat, M. A. (2019a). Neural and Statistical Machine Translation: A comparative error analysis, 393-403; Haji Sismat, M. A. (2019b). Inverse Translation Quality: A comparative analysis between human translation and post-editing, 2, 91-105.

	➢ Incorrect case ➢ Punctuation ❖ Style: ➢ Awkward ➢ Tautology

The purpose of the error annotation is also to find out patterns of errors in both SMT and NMT systems. Not only that these results can be useful for researchers and developers to improve the MT engines' performance, but also for educational purposes and public use, as nowadays the use of MT engines are popular, particularly among students. However, they do not know how to utilise the MT engines' full potential. Therefore, the patterns of errors in the two NMT engines would be convenient as a guideline for general users in the post-editing process. Moreover, the results of the comparative analysis of the two MT engines would give a great impression particularly to those who are not well-informed on how they can benefit from using the current NMT engines and how much the MT engines have been developed in the last few years.

RESULTS

Figure 1 presents the number of errors in both SMT and NMT systems, indicating a decrease of 72.2% and 73.1% of errors in MB and GT respectively. The significant decrease in the number of errors shows that MT is stepping in the right direction. Even so, it is crucial to find out the pattern of errors in both translation directions (Arabic-English and English-Arabic), as different language pairs, and translation directions may provide different results[11].

[11] Toral, A., Sánchez-Cartagena, V.M. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions; Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art?, 108(1), 109-120.

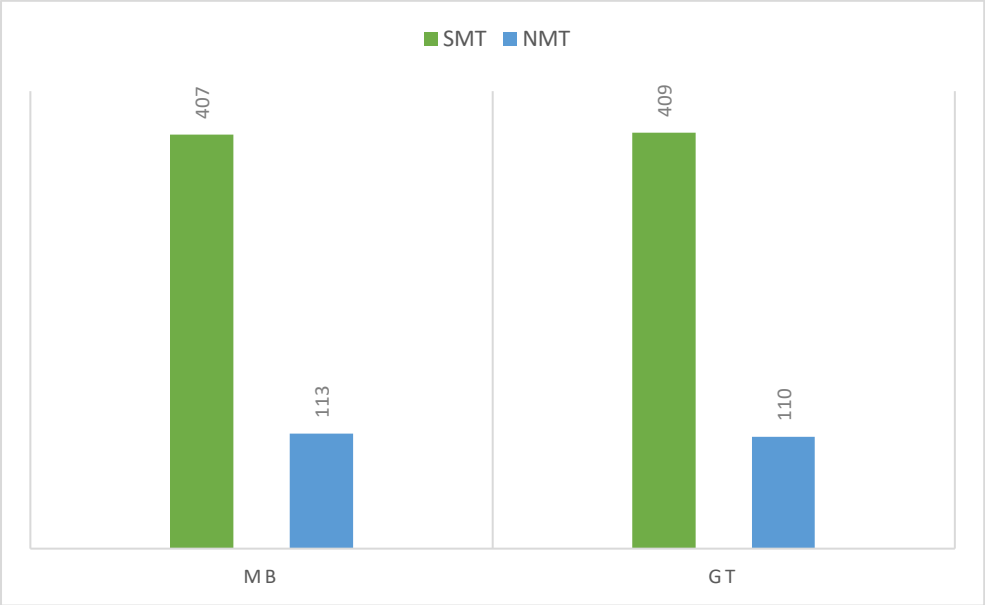

Figure 1: Number of errors in both SMT and NMT systems

The results in

Table 2 indicate that the major errors account for 8% and 16.4% of the total errors in GT and MB, respectively. The high percentage of minor errors indicates that the quality of the NMT outputs of both GT and MB can be improved if the post-editors thoroughly check and correct the minor errors. When compared to the data collected previously[12], the number of major and minor errors has decreased greatly, particularly that of hygiene-related errors in MB, indicating that the developers paid attention to the error analyses addressed previously.

Table 2

Number of major and minor errors in the two NMT engines

Type of error	**GT**		**MB**	
	Major	Minor	Major	Minor
Content-related	3	5	8	12
Grammar-related	1	56	1	34
Incorrect terms	2	17	8	22
Hygiene	1	17	-	18
Style	2	9	1	6
TOTAL	9 (8%)	104 (92%)	18 (16.4%)	92 (83.6%)

[12] Haji Sismat, M. A. (2019a). Neural and Statistical Machine Translation: A comparative error analysis, 393-403.

Arabic-English Translations

First, the present study looked at the number of errors in the Arabic-English translation in both types of MT systems and engines. Table 3 shows a consistent result indicating a decrease of 74.7% and 66.8% in the number of errors in both GT and MB, respectively. Also, the decrease in the number of errors in GT is considerably higher than that of MB.

Table 3

Number of errors in the AR-EN translation in both SMT and NMT systems

TEXT	**GT**		**MB**	
	SMT	NMT	SMT	NMT
AE1	45	5	38	5
AE2	54	17	61	18
AE3	84	22	62	25
AE4	70	20	68	28
TOTAL	253	64 (-74.7%)	229	76 (-66.8%)

Google Translate (GT). The results in Figure 2 indicate that initially incorrect term, distortion in meaning, awkward style, syntactic errors and too literal translations were mostly found in the Google Statistical Machine Translation (GSMT). However, when compared to the error analysis of the Google Neural Machine Translation (GNMT) outputs, the results reveal that distortion in meaning, awkward style, too literal translation and incorrect cases are no longer on the list of the most common types of errors, indicating only 1-3 errors each. In other words, the fluency of the Arabic-English translation outputs in GT has greatly improved. Even though incorrect terms and syntactic errors scored the highest, the present results show that the number of errors in the two types of errors has decreased by more than 50% when compared to the GSMT outputs.

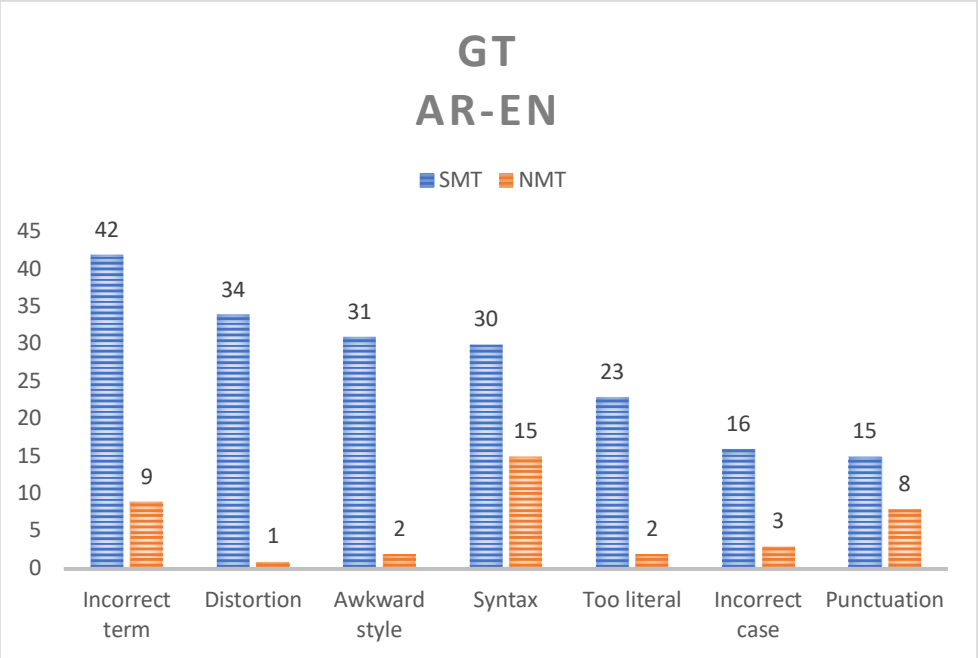

Figure 2: The types of errors commonly found in the AR-EN translation in both GSMT and GNMT

The results in Figure 2 also reveal that syntactic errors, incorrect terms, and incorrect cases contributed to the most errors in the Arabic-English translation outputs in GNMT. The other remaining types of errors seem almost 'resolved'. Therefore, the present study will only focus on the first three errors in this section:

- Syntactic errors:

Based on the results, syntactic errors scored the highest, accounting for 23.4% (15 out of 64 errors) in the current data. Even so, the number of syntactic errors dropped considerably, indicating that Google is heading in the right direction. Most of these errors are related to determiners, as Arabic sentences tend to be longer than English sentences. Typically, the Arabic sentences range from 20-30 words and may exceed 100 words as described by Al-Taani, Msallam, and Median[13]. Therefore, some determiners may be missing in the MT outputs.

[13] Al-Taani, A.T., Msallam, M.M., & Wedian, S.A. (2012), A top-down chart parser for analysing Arabic sentences. *The International Arab Journal of Information Technology*, 9(2), 109.

- Incorrect terms:

Based on the results, incorrect terms account for 14.1% of the errors in the Arabic-English translation outputs in GNMT. Even though overall the number of incorrect terms has decreased, unfortunately, it does not necessarily mean that it would provide the same results when translating different texts, depending on the MT inputs as seen in Table 4. The data revealed that AE4 contributed the most errors, indicating that GNMT may not be able to provide the same quality for every text type or content.

Table 4

The number of incorrect terms in the AR-EN translations in both GT and MB

TEXT	GT	MB
AE1	1	3
AE2	1	4
AE3	1	5
AE4	6	8
TOTAL	9	20

- Incorrect cases:

The results in Figure 2 only indicate eight errors in the Arabic-English Translation in GT. Even though it scored the third-highest, the number of errors seems relatively low. Based on the analysis, GT tends to use capital letters in the middle of the sentences, such as translating "وأيضا تقديم امتيازات خاصة..." as "… and also Offer special privileges", or cut the sentence using full stop, which subsequently, made the new sentence start with a capital letter. For example, "بما في ذلك عائدات الفنادق وعدد ليالي النزلاء...". The engine rendered the sentence as "…including hotel revenues. And the number of guest nights".

Microsoft Bing (MB). The results in Figure 3 reveal that incorrect terms, incorrect cases, distortions in meaning, awkward styles, and too literal translations contributed to the most errors in the Arabic-English translation outputs in Microsoft Bing Statistical Machine Translation (MBSMT). However, when compared to the error analysis of Microsoft Bing Neural Machine Translation (MBNMT), the number of awkward styles and too literal translations has decreased and no longer remains on top of the list.

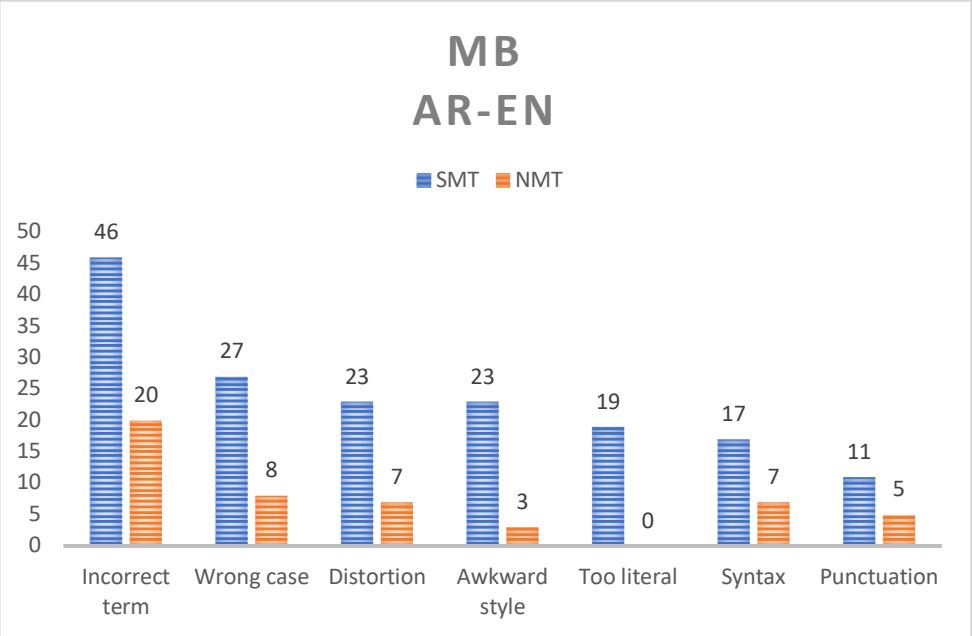

Figure 3: The types of errors commonly found in the AR-EN translation in both MBSMT and MBNMT

The above results also reveal that incorrect terms, incorrect cases, distortion and syntactic errors contributed the most errors in the Arabic-English MBNMT outputs. Even so, only the number of incorrect terms is fairly high when compared to the other types of errors. Therefore, the present study focuses on incorrect terms in this section:

- Incorrect terms:

The present analysis reveals that incorrect terms contributed the most errors, with 26.3% of the errors in the Arabic-English outputs. Also, the number of incorrect terms for each text is different. Based on the analysis, MBNMT failed to translate or transliterate proper nouns. For example, it failed to translate the word سبأ (Saba) when it is used in the different forms, such as السبئيين (Sabaeans) and العصر السبئي (Sabaean era). The MT engine rendered it as 'Sabes' and 'Seven Century'. It is also worth noting that the MT engine did not translate some words, such as مبارك and transliterated it as Mubarak. Again, the accuracy of the terms varies depending on the text types, as seen in Table 4.

English-Arabic Translations

The results in Table 5 indicate a significant decrease in the number of errors in the English-Arabic translations in both NMT engines, particularly MBNMT. Before reviewing the common types of errors, it is worth noting that the text types may be a contributing factor to the number of errors for each text. Texts EA1, EA4, and EA5 originated from the United Nations (UN)

legal documents, whereas the other two are economic texts. In the same table, the UN legal documents have fewer errors than the economic ones. A possible explanation for this is both GNMT and MBNMT initially used the UN legal documents to train their MT systems[14].

Table 5

Number of errors in the EN-AR translation in both SMT and NMT systems

TEXT	**GT**		**MB**	
	SMT	NMT	SMT	NMT
EA1	15	2	18	2
EA2	54	16	66	8
EA3	46	18	46	16
EA4	8	5	15	5
EA5	33	8	33	3
TOTAL	156	49 (-68.6%)	178	34 (-80.9%)

Google Translate (GT). The results in Figure 4 shows that incorrect terms, syntactic errors, distortions in meaning, too literal translations and awkward styles were commonly found in the English-Arabic GSMT outputs. However, when compared to the error analysis of the GNMT outputs, only syntactic errors remain problematic, accounting for 44.9% of the overall errors. Most of these syntactic errors are related to the conjunction "و" accounting for 17 out of 22 errors, indicating that these minor errors can be easily corrected and subsequently, the translation quality can be comprehensibly increased.

Previous data on GSMT addressed the word order issue in the English-Arabic direction. However, the present data reveals that most sentences are in the Verb-Subject-Object (VSO) order, which is preferable in formal writing.

[14] Haji Sismat, M. A. (2016). *Quality and productivity: A comparative analysis of human translation and post-editing with Malay learners of Arabic and English* (Doctoral dissertation).

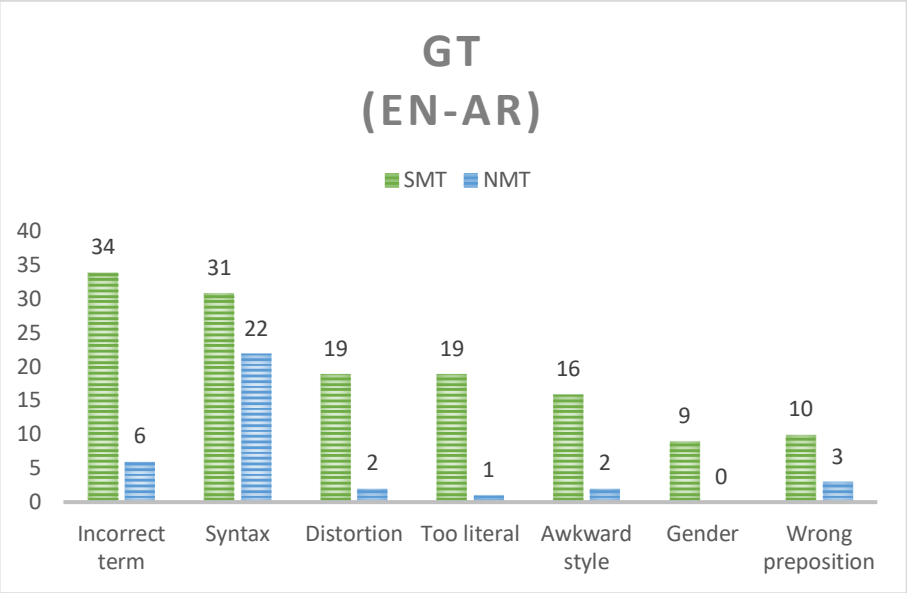

Figure 4: The types of errors commonly found in the EN-AR translation in both GSMT and GNMT

Microsoft Bing (MB). The results in Figure 5 show that syntactic errors, incorrect terms, awkward styles, distortions in meaning and too literal translations were commonly found in the English-Arabic MBSMT outputs. The present data reveals that incorrect terms, syntactic errors, and wrong preposition scored the highest. However, the number of these errors are fairly low, indicating that the accuracy and fluency of the English-Arabic translations have greatly improved.

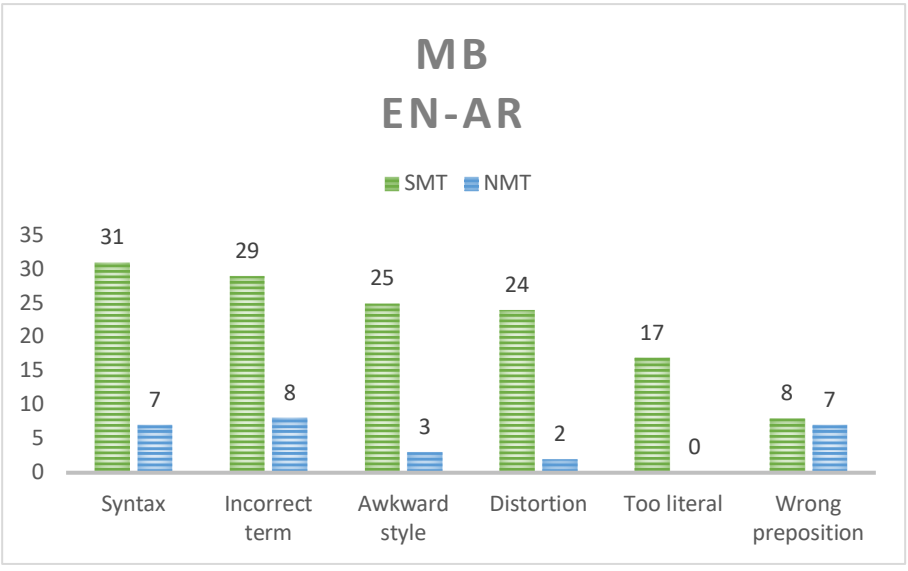

Figure 5: The types of errors commonly found in the EN-AR translation in both MBSMT and MBNMT

It may be worth noting that previous study[15] reported a high number of three types of misspellings: the letter Hamza "ء" Qat', the letter Ta' Marbuta "ة", and the letter Alif Maqsura "ى". However, the present data shows that these errors had been fixed by MB developers as none of these errors were spotted. It may also be worth noting that both MT engines inconsistently transliterated the Arabic names into English, which may affect the cohesiveness of the whole text.

DISCUSSION OF FINDINGS

The present study attempts to investigate the patterns of errors in NMT and SMT of Arabic and English. Based on the results of the error analysis, the present study answers each research question accordingly as follows:

1. Has the quality of the two MT engines (GT and MB) increased using the neural-approach when compared to the previous data collected from SMT?

   Based on the results, the error frequencies in both GT and MB have considerably decreased in both translation directions. The decrease indicates that both MT engines have managed to improve the quality of the translations comprehensibly, as the accuracy and fluency of the NMT outputs are noticeably better than that of the previous SMT outputs. The results also show that most of the errors are only minor errors that can be simply corrected if revised thoroughly.

2. What are the patterns of errors that currently exist in both GT and MB? Have the patterns of errors changed in both GT and MB?

   Due to a significant decrease in the number of errors in both MT engines, the patterns of errors have also changed. Based on the results, both syntactic errors and incorrect terms are the most common in both GT and MB. However, the former tends to be the highest in GT while the latter tends to be highest in MB. Despite having the most errors in respective MT engines, most of these errors decreased by more than 50% when compared to the previous data from SMT[16].

   In the Arabic-English translations, the results indicate that MB has more omissions than GT, indicating that users need to be aware of it as it may lead to distortion in meaning and subsequently, affect the translation quality. Syntactically, missing determiners are commonly found in MB when linking the sentences, which may be an issue that the

[15] Haji Sismat, M. A. (2019a). Neural and Statistical Machine Translation: A comparative error analysis, 393-403.
[16] Haji Sismat, M. A. (2016). *Quality and productivity: A comparative analysis of human translation and post-editing with Malay learners of Arabic and English* (Doctoral dissertation).

developers should look into, as it helps the readers see and understand the connection between ideas. However, in general, sentence structure in MB translations are better than that of GT. It is also worth noting that both NMT engines tend to mistranslate nouns, particularly Arabic names when transliterating them into English.

In the English-Arabic translations, it is worth noting that the errors relating to conjunctions are more commonly found in GT when compared to MB, indicating that MB looked into this matter as there were only four occurrences in the analysis.

The text types may also be a determining factor in the end-products, depending on the input of the MT engines. For example, the results of translating non-technical texts or sentences using the two engines may be poor as these engines mainly contain technical inputs. The study also revealed that there is a high possibility that good enough translations can be achieved when translating political and legal texts, primarily United Nations-related texts.

CONCLUSION

The present study has discussed the patterns of errors in both NMT and SMT systems, which may be used for further research and development among researchers and developers. The overall findings suggest that the quality of the NMT outputs has improved significantly in both English-Arabic and Arabic-English translations. However, the question of whether or not the NMT outputs are ready to be used depends entirely on the purpose of the post-editing tasks, the post-editor's level of experience and familiarity with the MT engines.

Most errors have been reduced in the NMT systems, including incorrect terms and syntactic errors which were reported as problematic in the previous SMT data. It is also worth noting that minor errors account for 92% and 83.6% of the total errors in both GT and MB. Therefore, if the post-editors are made aware of these errors and correct them thoroughly, their translations can be at least of good quality. It would be interesting to see whether or not the existing MT errors from the present study can be easily corrected by users, such as translation students and professional translators.

REFERENCES

Al-Samawi, A. M. (2014). Language errors in machine translation of encyclopaedic texts from English into Arabic: the case of Google Translate. *Arab World English Journal*, 182-211. Retrieved from https://awej.org/images/AllIssues/Specialissues/Translation3/17.pdf

Al-Taani, A.T., Msallam, M.M., & Wedian, S.A. (2012), A top-down chart parser for analysing Arabic sentences. *The International Arab Journal of Information Technology*, 9(2), 109-116. Retrieved from https://eis.hu.edu.jo/Deanshipfiles/pub109914508.pdf

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations 2015,* (pp. 1-15). San Diego, California.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., ... Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers,* 131-198. Retrieved from https://www.aclweb.org/anthology/W16-2301

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics,* 108(1), 109-120. Retrieved from https://www.degruyter.com/downloadpdf/j/pralin.2017.108.issue-1/pralin-2017-0013/pralin-2017-0013.xml

Daems, J., Macken, L., & Vandepitte, S. (2013). Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+PE. In *Proceedings of MT Summit XIV Workshop on Post-Editing Technology and Practice,* 2, 63-71. Retrieved from https://pdfs.semanticscholar.org/d845/3786f35a746fffcda098a1702f5f0b9759a2.pdf

Guerberof, A.A. (2012). *Productivity and quality in the post-editing of outputs from translation memories and machine translation* (Doctoral dissertation). Universitat Rovira I Virgili, Tarragona, Spain.

Haji Sismat, M. A. (2016). *Quality and productivity: A comparative analysis of human translation and post-editing with Malay learners of Arabic and English* (Doctoral dissertation). University of Leeds, Leeds, United Kingdom.

Haji Sismat, M. A. (2019a). Neural and Statistical Machine Translation: A comparative error analysis. In *Proceedings of 17th International Conference of Translation,* 393-403.

Haji Sismat, M. A. (2019b). Inverse Translation Quality: A comparative analysis between human translation and post-editing. *Journal of Arabic Linguistics and Literature*, 2, 91-105. Retrieved from http://www.unissa.edu.bn/journal/index.php/jall/article/view/113

Izwaini, S. (2006). Problems of Arabic machine translation: evaluation of three systems. *The British Computer Society (BSC),* London, 118-148. Retrieved from http://mt-archive.info/BCS-2006-Izwaini.pdf

Koponen, M., Aziz, W., Ramos, L., & Specia, L. (2012). Post-editing time as a measure of cognitive effort. In *Proceedings of WPTP,* 11-20. Retrieved from http://www.mt-archive.info/AMTA-2012-Koponen.pdf

Koponen, M., & Salmi, L. (2015). On the correctness of machine translation: A machine translation post-editing task. *The Journal of Specialised Translation,* 23, 118-136. Retrieved from http://www.jostrans.org/issue23/art_koponen.pdf

MeLLANGE. (2007). MeLLANGE: Multilingual eLearning in LANGuage Engineering. Retrieved from http://corpus.leeds.ac.uk/mellange/ltc.html

Papipeni, 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 on association for computational linguistics,* 311-318. Retrieved from https://aclanthology.info/pdf/P/P02/P02-1040.pdf

Secară, A. (2005). Translation evaluation: A state of the art survey. In *Proceedings of the eCoLoRe/MeLLANGE workshop*, 39-44. Retrieved from https://pdfs.semanticscholar.org/e5b3/a34db96b2e4ebb4d621bc4f6b8a9735e8f68.pdf

Specia, L, & Farzindar, A. (2010). Estimating machine translation post-editing effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10).* 33-41. Retrieved from https://pdfs.semanticscholar.org/6410/e3bf9c780bef4ada5a8eaac7532c9297d082.pdf

Toral, A., Sánchez-Cartagena, V.M. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. Retrieved from https://pdfs.semanticscholar.org/7b77/61e0c3c35278a8104994d8bd63fb0b91bb86.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems,* 1-11. Retrieved from https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., ... & Oflazer, K. (2014). Large scale arabic error annotation: Guidelines and framework. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference,* 2362-2369. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/956_Paper.pdf