Arabic and English Relative Clauses and Machine Translation Challenges

Khalil A. Nagi [1, *]

Received: 4 September 2023
Revised: 6 September 2023
Accepted: 25 September 2023

[1] Assistant Professor, Department of English – Faculty of Education and Sciences - University of Saba Region, Yemen.
* Corresponding author. E-mail: khalil.najji@usr.ac

https://doi.org/10.20428/jss.v29i3.2180

Arabic and English Relative Clauses and Machine Translation Challenges

Abstract:

The study aims at performing an error analysis as well as providing an evaluation of the quality of neural machine translation (NMT) represented by Google Translate when translating relative clauses. The study uses two test suites are composed of sentences that contain relative clauses. The first test suite composes of 108 pair sentences that are translated from English to Arabic whereas the second composes of 72 Arabic sentences that are translated into English. Errors annotation is performed by 6 professional annotators. The study presents a list of the annotated errors divided into accuracy and fluency errors that occur based on MQM. Manual evaluation is also performed by the six professionals along with a BLEU automatic evaluation using the Tilde Me platform. The results show that fluency errors are more frequent than accuracy errors. They also show that the frequency of errors and MT quality when translating from English into Arabic is lower than the frequency of errors and MT quality when translating from Arabic into English is also presented. Based on the performed error analysis and both manual and automatic evaluation, it is pointed out that the gap between MT and professional human translation is still large.

Keywords: fluency, accuracy, error analysis, evaluation, test suite.

عبارات الوصل باللغتين العربية والإنجليزية وتحديات الترجمة الآلية

الملخص:

تهدف هذه الدراسة إلى تحليل أخطاء الترجمة الآلية العصبية (NMT) وتقييم جودتها ممثلة بمترجم جوجل عند ترجمة عبارات الوصل. تستخدم الدراسة اختبارين مكونين من جمل تحتوي على عبارات وصل حيث يتكون الاختبار الأول من 108 جملة تمت ترجمتها من الإنجليزية إلى العربية بينما يتكون الاختبار الثاني من 72 جملة عربية تمت ترجمتها إلى الإنجليزية. تم تحديد الأخطاء بواسطة 6 خبراء. تقدم الدراسة قائمة مصنفة بأخطاء الترجمة بناءً على ما يعرف بمقياس الجودة متعددة الأبعاد (MQM). تم أيضًا إجراء تقييم بواسطة الخبراء الستة بالإضافة إلى تقييم BLEU الآلي. أظهرت النتائج أن أخطاء الطلاقة أكثر تكرارا من أخطاء الدقة وأن تكرار الأخطاء وجودة الترجمة الآلية عند الترجمة من الإنجليزية إلى العربية أقل من تكرار الأخطاء وجودة الترجمة الآلية عند الترجمة من العربية إلى الإنجليزية. توصلت الدراسة إلى أن الفجوة بين الترجمة الآلية والترجمة البشرية الاحترافية لا تزال كبيرة.

الكلمات الافتتاحية: الطلاقة، الدقة، تحليل الأخطاء، التقييم، الاختبار.

Introduction:

It is undeniable that the recent development of machine translation (MT) is remarkable. Machine translation has received a great interest recently and it has developed greatly in the past years. After the emergence of the Neural Machine Translation (NMT) systems, which have been considered a great breakthrough in the field of MT, research work has been done to evaluate it and compare the quality of translation produced by NMT systems to the quality of MT provided by the preceding systems, such as Phrase-Based Machine Translation (PBMT) systems and the Statistical Machine Translation (SMT) systems. In study that has been performed to compare the quality of both NMT and PBMT, it has been indicated that NMT outperforms PBMT in many aspects (Bentivogli et al., 2016; Toral and Sanchez-Cartagena; 2017; Klubicka et al., 2017; Popović, 2017, among others). It is pointed out that NMT degrades faster with sentence length as indicated by Bentivogli et al. (2016) and Koehn and Knowles (2017). However, Popović (2018) has stated that PBMT has no advantage over NMT in that aspect. Recent research in the field has also indicated that NMT systems have outperformed SMT systems (Sennrich and Zhang, 2019; Ahmadnia and Dorr, 2020; Saunders, 2022, among others). In regard to Arabic MT, Oudah et al. (2019), Maruf et al. (2019) and Diab (2021) have also reached similar conclusions that NMT is more fluent and has fewer issues.

However, the question regarding quality of MT is still unconcluded. Some studies have proposed that Machine translation has developed greatly and it is very close to human translation. Isabelle et al. (2017) have stated that neural machine translation (NMT) has developed greatly and it is very close to human translation when handling close language pairs such as English and French or English and Spanish. In the case of translating English into German and French, Levin et al. (2017) concluded that the fluency of NMT closes to human translation. It is also stated that the machine translation is in par or outperformed human professional translation in specific cases (Hassan et al., 2018; Popel et al., 2020). However, despite the great progress of machine translation, evidence has been presented that the gap between human and machine translation is still big and that machine translation has not achieved human parity (Toral et al., 2018; Freitag et al., 2021).

Recent analyses of MT errors show that MT is still riddled with errors and they propose that more and more effort must be spent on identifying the specific nature of errors. The importance of fine-grained studies to the development of MT is evident. That is because they provide a clear insight into the points of weakness and strength of MT systems by pointing out detailed analysis of error typology which helps in the development of the MT systems as well as in the facilitation of the post-editing process (See Daems et al., 2014; Popovic, 2021; Kocmi et al., 2022; Rivera-Trigueros, 2022, among others).

Motivation and Related Work:

Despite its great development, the quality of machine translation output is still a matter of debate, especially when it comes to low-source languages like Arabic. Recent surveys indicate that the state achieved by Arabic MT systems have not achieved the same level of quality compared to other languages and that more improvement is required (Ameur et al., 2020; Zakraoui et al., 2021; Darwish et al., 2021). Zakraoui et al. (2021) have performed a survey on the challenges of Arabic MT. they indicate that main Arabic MT challenges are both linguistic and technical. The linguistic issues come from the morphological richness and syntactic nature of Arabic which makes it divergent from languages like English. Such divergence unsurprisingly raises many MT issues. The study has shown the research on Arabic NMT has increased recently and that some efforts have been done to evaluate the effectiveness of MT. It also indicated that many challenges in various aspects, including accuracy and fluency, need to be addressed. Ameur et al. (2020) have also performed a survey on the general topics of research studies developed in Arabic MT. According to them, the main focus has been on translating Arabic into English. Translating English into Arabic has been of secondary significance which is really a big deal since it seems that more challenges will appear when investigating the challenges of English-to-Arabic MT. They have also indicated that syntactic word reordering has been heavily studied and that is in term of free order. However, the focus of the study will be mainly on English to Arabic and the freedom of word order as indicated earlier will not be the issue since relative clauses in Arabic force VS word order. Ameur et al. (2020) concluded that there are

still a lot of Arabic-related linguistic problems that need a lot of investigation. That is because they cause significant challenges for MT. That is why the current study will participate greatly in this field.

In addition, relative clauses have been and still are a great and an interesting issue in the field of MT. They have been considered a problem for designing controlled language for MT (Mitamura, 1999; Cardey et al., 2004; T Aikawa, 2007, among others). In recent translation related research, relative clauses have also been a significant part of various test suites or challenges test sets used to investigate the MT quality (Isabelle et al., 2017; Isabelle & Kuhn, 2018; Avramidis & Macketanz, 2022). In addition, certain procedures have been proposed in the literature to deal with the difficulties they form for MT system such as source text simplification (Hasler et al., 2017; Štajner & Popović, 2018; Sulem et al., 2020). The interest in relative clauses in the MT field is natural due to the interesting linguistic nature of the structure. Due to their distinctive syntactic nature, relative clauses have been a heated topic of discussion in the field of general linguistics for a very long time and they still are.

The matter is more interesting in the field of MT when languages like English and Arabic are involved. Relative clauses in the two languages exhibit fascinating syntactic convergences. They also show great morphological convergences which is expected due to the fact that English has a comparatively poor morphology and Arabic is known for its very rich morphology.

It is known that Arabic allows free variation in terms of word order of simple clauses where a simple Arabic clause can have either an SV or a VS word order. However, such variation is limited when it comes to Arabic relative clauses where the VS structure becomes obligatory. On the other hand, English clauses, including relative clauses, follows the SV structure. This makes the two languages show completely different word orders in relative clauses. It is also known that, unlike English, Arabic is a morphologically rich language. Arabic exhibits a much greatly richer inflectional system in terms of number, gender and case. The agreement in relative clauses is not limited to the subject-verb agreement or the noun-adjective agreement. Relative pronouns in Arabic agree in number and person with the relative head. They also agree with the

verb in the case of subject relatives and with presumptive pronouns in various cases. (For more on relative clauses in Arabic, check Mohammad, 1990; Aoun et al., 1994; Soltan, 2007; Aoun et al., 2010).

The two languages also vary greatly in terms of presumptive pronouns in relative clauses. Presumptive pronouns in the context of relative clauses refer to the "pronominal elements which occur in the relativization position and they have the same reference as the relative heads" (Nagi, 2016, p. 39). Depending on the context in which they are used, presumptive pronouns are classified into two types: *intrusive presumptive pronouns and true presumptive pronouns*. English makes use of what is known as intrusive resumption where presumptive pronouns are used in contexts where movement is not allowed as a device to save the grammaticality of the structure. Such contexts are known syntactically as island contexts. Arabic also uses intrusive presumptive pronouns. However, it also used what is known as true presumptive. That is to say, a presumptive pronoun is used productively and it appears where a gap is supposed to appear. Such type of presumptive pronoun is not used in English but it is allowed in Arabic as shown in examples (1 & 2) below.

1. I met the girl that Bills loves (*her).

2. qabaltu al-binta allati yuhibu-ha bill.

met.1sg.past the-girl that love.3sg.pres-her Bill

'I met the girl that Bills loves.'

In the above example, the occurrence of the pronoun her is problematic in English. However, the occurrence of the presumptive pronoun *-ha* (her) in a correspondent structure in Arabic is regular. (For more on relative clauses and presumptive pronouns, see Shlonsky,1992; McKee & McDaniel, 2001; Ouhalla, 2004; Aoun et al., 2010; Nagi, 2016, 2022, among others.)

Another interesting aspect in the context of MT is regarding control and exceptional case marking (ECM) structures. It is known that English control and ECM verbs take infinitive complements which, therefore, do not show any agreement marking. However, verb in the correspondent Arabic structures agree with the matrix subjects

or objects. In the case of relative clauses, such agreement mostly involves the relative pronouns as well.

Based on all that, the investigation of the MT quality is a very appealing topic when relative clauses in English and Arabic with all their convergences are involved. It should be noted that, a single study in Arabic MT was devoted to relative clauses despite the remarkable nature of the structure. The study was conducted by Tratz et al. (2014) and its focus was limited to the challenges related to presumptive pronouns in Arabic relative clauses when translated into English. The current study, however, provides a fine-grained error analysis that includes all accuracy and fluency issues in the Arabic MT of English relative clauses and the English MT of Arabic relative clauses. In addition, it provides both manual and automatic evaluations of the translation. Test suites are constructed in order to perform the required error analysis and evaluation of the produced MT and it is needless to say that a fine-grained analysis of error is crucial to the development of MT system and the facilitation of the post-editing process, as mentioned earlier.

Methodology:

The Test Suites

Test suites are recommended and assumed effective and straightforward in evaluating how MT systems deal with hard and specific translation problems (Hardmeier, 2015; Guillou et al., 2018). To achieve the objectives of this study, two test suites with a total of 180 sentence pairs are constructed to provide a general evaluation of translation quality as well as an analysis of the translation issues that Google Translate encounters when translating English and Arabic relative clauses.

The first test suite used in the study is composed of 108 English sentences that contain relative clauses. The other test, however, is composed of 72 Arabic sentences that are structured in the same way and translated into English. Google Translate is used to translate these sentences. The sentences are collected from various online news resources. The news articles also are of different types such as political , economic, entertainment , and sport, etc.

The divergences that relative clauses in English and Arabic exhibit ensure that the chosen sentences from one language are divergent from their equivalent in the other language. The test suites, therefore, are suitable to evaluate the MT capability of dealing with the convergences that English and Arabic relative clauses exhibit.

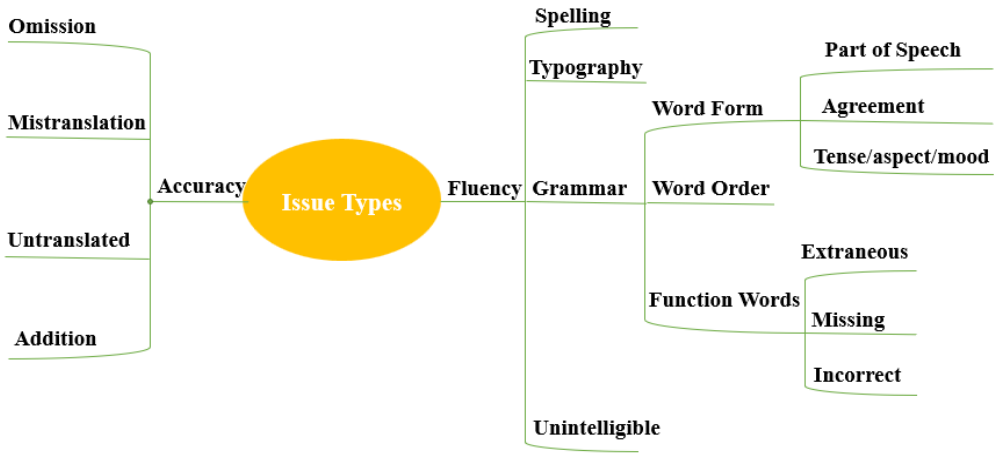
It should be noted that there are certain syntactic phenomena such as pied-piping and preposition stranding are not allowed in Arabic. Arabic varies from English and it uses a presumptive pronoun in the relativization site after the preposition for such cases. In addition to this, verbs in Arabic control and ECM structures show agreement. This leads to the use of fewer sentences in the second test suite to avoid unnecessary repetition of the issues.

Error Analysis and Evaluation

The annotation of errors is manually performed by six professional annotators who are native speakers of Arabic, fluent in English, and have a long experience in the fields of annotation and translation. The annotators are required to identify and label the errors. The classification of the annotated errors is based on the Multidimensional Quality Metrics (MQM) typology (https://themqm.org/). MQM, as identified by Lommel et al. (2014), is flexible framework that is primarily used to evaluate MT and to deal with the shortcomings of the previous systems used for MT quality evaluation. The typology provided by MQM classified translation errors into seven main dimensions: terminology, accuracy (adequacy), linguistic conventions (fluency), style, locale conventions, audience appropriateness, and design and markup. Such dimensions are defined and classified further. The main issues related to accuracy are mistranslation, addition and omission, whereas the main issues related to linguistic conventions are grammar, punctuation and spelling.

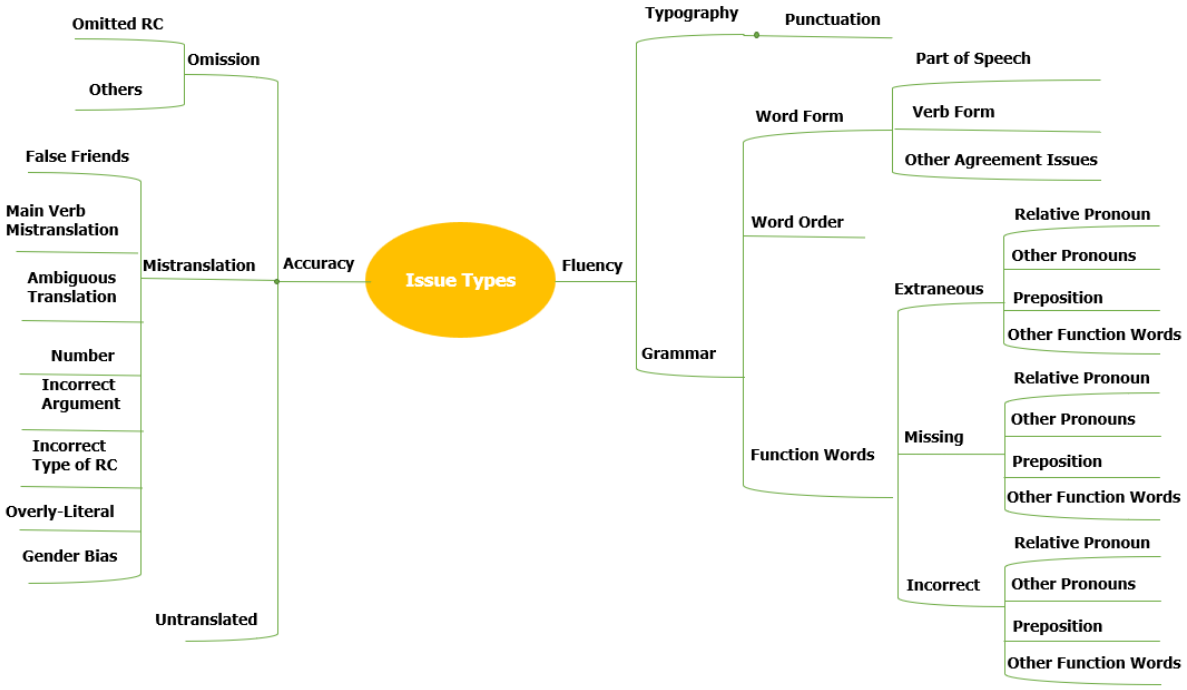
The MQM framework is chosen in this study due to the fact that it is well-established and flexible. However, it proposes a very broad tag-set which makes it impractical to perform an annotation of a specific purpose. Therefore, due to the nature of the study, accuracy and fluency errors are annotated, to which the core error types as proposed by the MQM are presented in Figure 1 below (as proposed in Klubička et al., 2017).

Figure 1. The MQM core error types



In this study, however, since relative clauses are involved and due to the morphological and syntactic convergences between English and Arabic, the annotated set of errors under accuracy and fluency are customized as represented in the Figure 2 below.

Figure 2. The English & Arabic annotated error types



The annotated errors fall within the two main categories: fluency errors and adequacy errors as shown in Figure 2. It should be noted here that, due to the structure under study, function words are classified further, and issues such as incorrect argument,

incorrect type of relative clauses, missing relative clauses, and main verb mistranslation are included. The customized issues are explained as follows.

Fluency Issues

Verb Form: This issue concerns the general form of the verb which includes its agreement with subject, tense, aspect, etc. The issues related to the verb form are labelled separately due to the assumption that relative clauses work as a distractor that causes an issue for main verb and affects its form.

Other Agreement Issues: This includes all other agreement issues that are not related to verb form, such as noun-adjective agreement, case, etc.

Function Words: Function words here are divided into four categories. That is relative pronouns, other pronouns (pronouns other than relative pronouns such as personal pronouns, reflexives, demonstratives, etc.), prepositions, and other function words (function words other than pronouns which mostly include determiners and conjunctions). This is due to the fact that English and Arabic varies in the form and agreement of the relative element, as well as the use of presumptive pronouns as indicated in the study earlier. Prepositions are also separated due to the fact that pied-piping usually involves a preposition and there is a great variation between English and Arabic as mentioned before. According to Cable (2012), pied-piping refers to those structures where a constituent larger than expected moves to a higher position. In the case of relative clauses, the movement, therefore, involves more than the relative pronoun. Check the pied-piped phrases in bold in the examples below.

3. She was a devoted mother to her daughter ***to whom*** she was deeply attached.
4. Shapiro started taking meetings all over town, ***a couple of which*** Maddow joined.

Punctuation: Punctuation here refers to the availability or non-availability of commas needed to separate the relative clause from the rest of the sentence. It should be mentioned here that the use of commas, in general, is different between the two languages. The issue is more serious since commas, for example, are used to separate

non-restrictive relative clauses from the rest of the sentence in English which is not in Arabic.

Accuracy Issues

Incorrect argument: This refers to cases in which an argument in a relative clause occurs in place of another argument, i.e., a subject occurs as an object or the other way around.

Incorrect type of relative clauses: This refers to cases where, for instance, a headless relative clause is translated as a headed one.

Missing relative clauses: This refers to cases where the whole relative clause is omitted from the translation.

Main verb mistranslation: This refers to cases where the main verb is mistranslated as a part of the relative clause.

Number: This issue refers to the translation error in which a dual in Arabic is simply translated into plural in English without putting any indication that the matter involves, say, two entities or two people.

In addition to error analysis, both manual evaluation and automatic evaluation of the produced MT are also conducted in this study. Despite the fact that manual evaluation is expensive and time consuming, it is performed in the study due to the nature of the source text involved. The study here deals also with inter-sentential issues since relative clauses are long-distance dependencies. It is proposed in the literature that long-distance dependencies are not just problematic for MT systems, but it is also proposed that automatic evaluation performs poorly in the case of anaphoric pronouns (Guillou & Hardmeier, 2018; Guillou et al., 2018). Therefore, a manual evaluation of MT is performed here to ensure the integrity of the results.

The manual evaluation is conducted by the six annotators using a 7-point Likert scale. The scale ranges from 0 to 6 where 0 means no meaning preserved and 6 means perfect meaning and grammar. An automatic evaluation using BLEU is also performed.

Results

Error Analysis

1. English to Arabic Errors

An error typology along with the number and percentage of annotated errors of the English to Arabic translated sentences is presented in Table 1 below. Table 1 shows that the total number of fluency errors is (184) where verb form comes at the top with (36) errors followed by missing relative pronouns with (27) errors. The total number of accuracy errors is (78) with false friends comes at the top with 35 errors followed by Missing content words with (13) errors and incorrect argument with (9) errors. The fluency errors form 70.23% of the total errors while the accuracy errors form 29.77% of the total errors.

Table 1. Number of Errors in Arabic Sentences Translated from English

Dimensions	Types of Errors	No. of Errors
Fluency	Incorrect Relative Pronouns	8
	Incorrect Pronouns	22
	Incorrect Prepositions	2
	Incorrect Function Words	2
	Missing Relative Pronouns	27
	Missing Pronouns	11
	Missing Prepositions	15
	Missing Function Words	6
	Extraneous Relative Pronouns	1
	Extraneous Pronouns	6
	Extraneous Prepositions	1
	Extraneous Function Words	4
	Verb Form	36
	Other Agreement Issues	6
	Part of Speech	0
	Word Order	20
	Punctuation	17
Total Fluency Errors		**184**
Accuracy	Missing Content Words	13
	Omitted Relative clauses	1
	False Friends	35
	Main Verb Mistranslation	1
	Ambiguous Translation	4
	Number	0
	Incorrect Argument	9
	Incorrect Relative Clause Type	4
	Overly-Literal	1
	Gender Bias	5
	Untranslated	5
Total Accuracy Errors		**78**
The Total Errors		**262**

2. Arabic to English Errors

Table 2 below presents the number and the percentage of the annotated errors in the English sentences translated from Arabic. Table 2 shows that the total number of fluency errors is (54) where punctuation comes at the top with (10) errors followed by verb form and word order with (9) errors of each type. The total number of accuracy errors is (34) with false friends comes at the top with 12 errors followed by Missing content words with (7) errors and incorrect argument with (6) errors. The fluency errors form 61.36% of the total errors while the accuracy errors form 38.64% of the total errors.

Table 2. Number of Errors in English Sentences Translated from Arabic

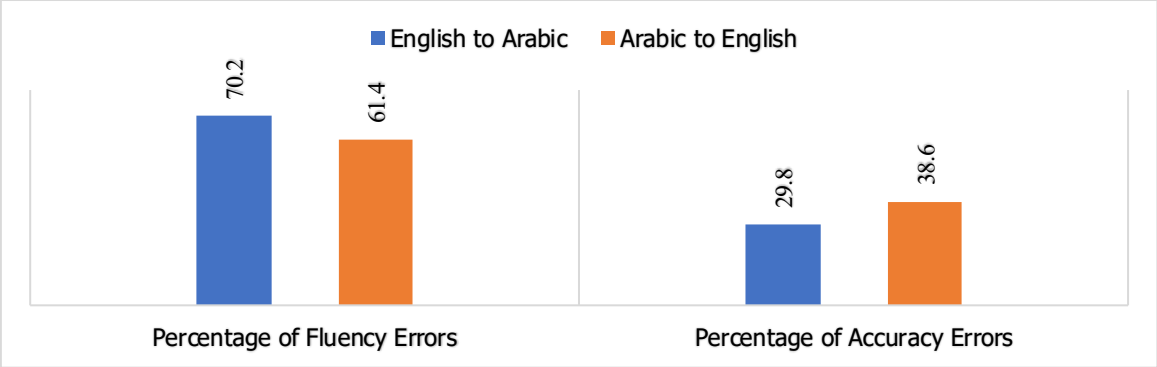
Dimensions	Types of Errors	No. of Errors
Fluency	Incorrect Relative Pronouns	2
	Incorrect Pronouns	4
	Incorrect Prepositions	3
	Incorrect Function Words	0
	Missing Relative Pronouns	4
	Missing Pronouns	2
	Missing Prepositions	0
	Missing Function Words	5
	Extraneous Relative Pronouns	1
	Extraneous Pronouns	2
	Extraneous Prepositions	1
	Extraneous Function Words	1
	Verb Form	9
	Other Agreement Issues	0
	Part of Speech	1
	Word Order	9
	Punctuation	10
Total Fluency Errors		**54**
Accuracy	Missing Content Words	7
	Omitted Relative clauses	0
	False Friends	12
	Main Verb Mistranslation	0
	Ambiguous Translation	2
	Number	2
	Incorrect Argument	6
	Incorrect Relative Clause Type	0
	Overly-Literal	1
	Gender Bias	3
	Untranslated	1
Total Accuracy Errors		**34**
The Total Errors		**88**

3. Fluency vs Accuracy Errors

Checking the data above, it can be grasped that the frequency of errors in the English to Arabic MT are higher than the frequency of errors in the Arabic to English MT. This is a plausible outcome since it involves two languages that vary in the richness of

morphology. The data also show that the frequency of fluency errors is higher than that of accuracy errors in both English and Arabic MT as shown in Figure 3.

Figure 3. Frequency of fluency and accuracy errors in English and Arabic MT



Evaluation

1. Manual Evaluation

Manual evaluation is performed on the produced MT. As mentioned earlier, a 7-point Likert scale was used ranging from 0 which means no meaning preserved to 6 which means perfect meaning and grammar. The evaluation is done by the 6 professional annotators and the results were as shown in Table 3 below.

Table 3. Manual evaluation of MT

Language Pair	Mean	Standard Deviation
En > Ar	4.2	0.31
Ar > En	4.8	0.34

Table 3 above presents the mean and standard deviation which indicates that the translation from English to Arabic retains most of the meaning of the source and that it has some grammar mistakes. It also indicates that the meaning of the translation is almost consistent with the source with few grammar mistakes in the translation from Arabic to English. It is also clear from the table above that, based on the manual evaluation, the quality of the Arabic to English MT is higher than the quality of the English to Arabic MT.

2. Automatic Evaluation

One of the most common automatic evaluation methods used to evaluate the quality of machine translation is BLEU. Papineni et al. (2002) introduced this method to avoid the cost and time consumption of human evaluation. When the evaluation is performed, an MT is compared with one or more reference translation. BLEU counts

the number of the matching words and compares the n-grams of the MT with that of the reference translations.

The automatic evaluation in this study is limited to BLEU. An evaluation is performed on the translation here using the Tilde MT platform exhibiting the results in the following table.

Table 4. Automatic evaluation of MT

Language Pair		1-gram	2-gram	3-gram	4-gram
En > Ar	Individual	97.22	92.35	88.41	84.77
	Cumulative	96.40	93.96	91.81	**89.81**
Ar > En	Individual	88.72	78.38	70.13	62.94
	Cumulative	88.34	83.03	78.38	**74.11**

Despite the fact that automatic evaluation shows that the translation has a higher quality than what the manual evaluation shows, the automatic evaluation also indicates that Google Translate provides a higher quality translation when translating the English sentences into Arabic.

Discussion

This study aimed mainly to identify the errors that occur when relative clauses are translated from English to Arabic and vice versa using Google Translate. From the results pointed out in the previous sections, it is clear that there are plenty of fluency and accuracy issues that MT exhibits when translating sentences that contain relative clauses. The matter is more serious when translating from English to Arabic. The error types are classified under fluency and accuracy (referred to as adequacy too). Despite the fact some errors are classified further here due to the nature of the study, many of the annotated errors appear in recent error analyses (Popović, 2021; Kocmi et al., 2022). Therefore, more in-depth studies are needed to identify errors in specific constructions and to point out their position and structural specifications.

The study also aimed to evaluate the generated MT and both professional manual evaluation and automatic evaluation were performed. The performed evaluations along with the annotation of errors indicate that the gap between professional human translation and MT is still not small. This goes with what Toral et al. (2018) and Freitag et al. (2021) proposed that MT has not achieved human parity as opposed to what was assumed by Hassan et al. (2018) and Popel et al. (2020). According to this study, the gap is bigger when translating from English to Arabic which is due to the poor morphology of English and the rich morphology of Arabic.

In regard to Arabic relative clauses translated into English, it is mentioned earlier that a study that Tratz et al (2015) performed a study that mainly aimed to investigate the challenges of MT when dealing with presumptive pronouns in translating Arabic to

English sentences. The system is still suffering when dealing with such issues. Different issues such as incorrect argument and ambiguous translation result due to the convergence between English and Arabic in such aspect. A future study will be performed to investigate such aspects.

Conclusion

The study came out with a list of accuracy and fluency issues that occur in the MT produced by Google Translate when translating English and Arabic sentences that contain relative clauses to the corresponding target language. The study classified the issues fluency and accuracy issues based on the MQM framework and presented a comparison which showed that the number of fluency issues is higher than the number of accuracy issues when translating English sentences into Arabic or the other way around. Moreover, the study showed that the frequency of both accuracy and fluency issues is higher when translating from English to Arabic. Both manual and automatic evaluations were performed. The results showed that the machine translation is still not on bar with professional human translation. They also exhibited that the translation from Arabic to English produced by Google Translate is higher quality than the translation from English into Arabic.

Acknowledgement

This study was funded by the Literature, Publishing and Translation Commission, Ministry of Culture, Kingdom of Saudi Arabia under [132/2023] as part of the Arabic Observatory of Translation.

References

Ahmadnia, B., & Dorr, B. J. (2020). Low-Resource multi-domain machine translation for Spanish-Farsi: Neural or Statistical?. *Procedia Computer Science*, *177*, 575-580. https://doi.org/10.1016/j.procs.2020.10.081

Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M., & Lozano, C. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proceedings of Machine Translation Summit XI: Papers*.

Ameur, M. S. H., Meziane, F., & Guessoum, A. (2020). Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Review*, *38*, 100305. https://doi.org/10.1016/j.cosrev.2020.100305

Aoun, J., Benmamoun, E., & Sportiche, D. (1994). Agreement, word order, and conjunction in some varieties of Arabic. *Linguistic inquiry*, 195-220. http://www.jstor.org/stable/4178858

Aoun, J., Choueiri, L., & Benmamoun, E. (2010). *The syntax of Arabic*. New York: Cambridge University Press.

Avramidis, E., & Macketanz, V. (2022). Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 514-529). https://aclanthology.org/2022.wmt-1.45/

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2016). Neural versus phrase-based machine translation quality: A case study. In *2016 Conference on Empirical Methods in Natural Language Processing* (pp. 257-267). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/d16-1025

Cable, S. (2012). Pied-piping: Introducing two recent approaches. *Language and Linguistics Compass*, *6*(12), 816-832. https://doi.org/10.1111/lnc3.12001

Cardey, S., Greenfield, P., & Wu, X. (2004,). Designing a controlled language for the machine translation of medical protocols: The case of English to Chinese. In Conference of the Association for Machine Translation in the Americas (pp. 37-47). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30194-3_5

Daems, J., Macken, L., & Vandepitte, S. (2014). On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship. In *LREC* (pp. 62-66).

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., ... & Mubarak, H. (2021). A panoramic survey of natural language processing in the Arab world. *Communications of the ACM, 64*(4), 72-81. https://doi.org/10.1145/3447735

Diab, N. (2021). Out of the BLEU: An error analysis of statistical and neural machine translation of WikiHow articles from English into Arabic. *CDELT Occasional Papers in the Development of English Education*, *75*(1), 181-211. 10.21608/OPDE.2021.208437

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460-1474. https://doi.org/10.1162/tacl_a_00437

Guillou, L., & Hardmeier, C. (2018). Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4797-4802). https://aclanthology.org/D18-1513

Guillou, L., Hardmeier, C., Lapshinova-Koltunski, E., & Loáiciga, S. (2018). A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 570-577). https://aclanthology.org/W18-6435

Hardmeier, C. (2015, September). On statistical machine translation and translation theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation* (pp. 168-172). https://aclanthology.org/W15-2522/

Hasler, E., de Gispert, A., Stahlberg, F., Waite, A., & Byrne, B. (2017). Source sentence simplification for statistical machine translation. *Computer Speech & Language*, *45*, 221-235. https://doi.org/10.1016/j.csl.2016.12.001

Isabelle, P., & Kuhn, R. (2018). A Challenge set for French--> English machine translation. *arXiv preprint arXiv:1806.02725*. https://doi.org/10.48550/arXiv.1806.02725

Isabelle, P., Cherry, C., & Foster, G. (2017). A Challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2486-2496). https://doi.org/10.48550/arXiv.1704.07431

Klubička, F., Toral Ruiz, A., & Sánchez-Cartagena, M. V. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 121-132. http://dx.doi.org/10.1515/pralin-2017-0014

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., ... & Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 1-45). https://aclanthology.org/2022.wmt-1.1

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. In *First Workshop on Neural Machine Translation* (pp. 28-39). Association for Computational Linguistics. https://aclanthology.org/W17-3204

Levin, P., Dhanuka, N., & Khalilov, M. (2017). Machine translation at booking. com: Journey and lessons learned. *arXiv preprint arXiv:1707.07911*. https://doi.org/10.48550/arXiv.1707.07911

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12), 0455-463. https://doi.org/10.5565/rev/tradumatica.77

McKee, C., & McDaniel, D. (2001). Resumptive pronouns in English relative clauses. *Language acquisition*, *9*(2), 113-156. https://doi.org/10.1207/S15327817LA0902_01

Mitamura, T. (1999). Controlled language for multilingual machine translation. In *Proceedings of Machine Translation Summit VII* (pp. 46-54).

Mohammad, M. (1990). The problem of subject-verb agreement in Arabic: Towards a solution. *Perspectives on Arabic linguistics I*, 95-127.

Nagi, K. A. (2016). Gap and resumption strategies in Modern Standard Arabic restrictive relative clauses: A minimalist approach. [Doctoral dissertation]. SRTM University, Nanded, India.

Nagi, K. A. (2022). Free relatives in Standard Arabic: An agree-based account. *University of Saba Region Scientific Journal, 5*(1). https://doi.org/10.54582/TSJ.2.2.53

Oudah, M., Almahairi, A., & Habash, N. (2019). The Impact of preprocessing on Arabic-English statistical and neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track* (pp. 214-221). European Association for Machine Translation. https://aclanthology.org/W19-6621

Ouhalla, J. (2004). Semitic relatives. in *Linguistic Inquiry, 35* (2), 288-300. https://doi.org/10.1162/002438904323019084

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). https://doi.org/10.3115/1073083.1073135

Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., & Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, *11*(1), 4381-4381. https://doi.org/10.1038/s41467-020-18073-9

Popović, M. (2017). Comparing language related issues for NMT and PBMT between German and English. In *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 209.

Popović, M. (2018). Language-related issues for NMT and PBMT for English–German and English–Serbian. *Machine Translation*, *32*, 237-253. https://doi.org/10.1007/s10590-018-9219-5

Popović, M. (2021). On nature and causes of observed MT errors. In *Proceedings of Machine Translation Summit XVIII: Research Track* (pp. 163-175). https://aclanthology.org/2021.mtsummit-research.14

Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, *56*(2), 593-619. https://doi.org/10.1007/s10579-021-09537-5

Saunders, D. (2022). Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, *75*, 351-424. https://doi.org/10.1613/jair.1.13566

Sennrich, R., & Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *57th Annual Meeting of the Association for Computational Linguistics* (pp. 211-221). Association for Computational Linguistics (ACL). https://aclanthology.org/P19-1021/

Shlonsky, U. (1992). Resumptive pronouns as a last resort. *Linguistic inquiry*, *23*(3), 443-468. https://www.jstor.org/stable/4178780

Soltan, U. (2007). *On formal feature licensing in minimalism: Aspects of Standard Arabic morphosyntax*. [Doctoral dissertation]. University of Maryland, College Park.

Štajner, S., & Popović, M. (2018). Improving machine translation of English relative clauses with automatic text simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, 39-48. https://aclanthology.org/W18-7006

Sulem, E., Abend, O., & Rappoport, A. (2020). Semantic structural decomposition for neural machine translation. In *Proceedings of the ninth joint conference on lexical and computational semantics* (pp. 50-57). https://aclanthology.org/2020.starsem-1.6

Toral, A., & Sánchez-Cartagena, V. M. (2017). A Multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1063-1073). https://aclanthology.org/E17-1100

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 113-123). https://aclanthology.org/W18-6312

Tratz, S., Voss, C., & Laoudi, J. (2014). Resumptive pronoun detection for Modern Standard Arabic to English MT. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)* (pp. 42-47). https://aclanthology.org/W14-1008

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic machine translation: A survey with challenges and future directions. *IEEE Access*, *9*, 161445-161468. https://doi.org/10.1109/ACCESS.2021.3132488