Fine-Grained Evaluation of English to Arabic Neural Machine Translation: A Case Study of Educational Research Abstracts

التقييم الدقيق للترجمة الآلية العصبية من الإنجليزية إلى العربية:
دراسة حالة للملخصات البحثية التربوية

https://aif-doi.org/AJHSS/119507

Hesham A. Almekhlafi[*]

Khalil A. Nagi[**]

*Amran University
almekhlafihesham@gmail.com
**University of Saba Region
khalil.nagi@usr.ac

Abstract:

The study aims to investigate the quality of neural machine translation when translating research paper abstracts from English to Arabic. It performs an error analysis and provides an evaluation of the quality of neural machine translation (NMT) represented by Google Translate and Microsoft Translator. The research team selects 25 English research paper abstracts in education from well-known Scopus scientific journals issued in English speaking countries. These abstracts are then translated into Arabic using both Google Translate and Microsoft Translator. The error analysis is based on the typology of errors introduced by Multidimensional Quality Metrics (MQM). A professional evaluation is also conducted using the Scalar Quality Metric evaluation (SQM) as proposed in Freitag (2021). The study finds that the translation outputs of academic texts like abstracts of education research papers are still not up to standards when translating English educational research abstracts into Arabic. There are various types of translation errors. However, there is a slight difference in translation quality and number of errors in favor of Google Translate compared to Microsoft Translator. However, it is included that NMT system still requires a lot of training, and more Arabic corpora need to be built.

Keywords: machine translation, evaluation, fine-grained, errors, English-Arabic, abstracts, education research.

التقييم الدقيق للترجمة الآلية العصبية من الإنجليزية إلى العربية: دراسة حالة للملخصات البحثية التربوية

د. هشام عبدالله عبده المخلافي
د. خليل عبدالسلام خالد ناجي

تهدف الدراسة إلى التحقيق في جودة الترجمة الآلية العصبية عند ترجمة ملخصات أوراق البحث العلمي من اللغة الإنجليزية إلى اللغة العربية، حيث تقوم الدراسة بتحليل الأخطاء وتقييم جودة الترجمة الآلية العصبية (NMT) الممثلة في Google Translate و Microsoft Translator. يقوم فريق البحث باختيار 25 ملخص ورقة بحثية إنجليزية في مجال التعليم من مجلات سكوبس العلمية المعروفة والصادرة في البلدان الناطقة باللغة الإنجليزية. ثم تتم ترجمة هذه الملخصات إلى العربية باستخدام كل من Google Translate و Microsoft Translator. تعتمد الدراسة في تحليل الأخطاء على تصنيف الأخطاء التي قدمتها معايير الجودة متعددة الأبعاد(MQM) . كما يتم إجراء تقييم احترافي باستخدام تقييم معيار الجودة القياسي (SQM) كما هو مقترح في (Freitag، 2021). تخلص الدراسة إلى أن مخرجات ترجمة النصوص الأكاديمية مثل ملخصات أوراق أبحاث التربية لا تزال غير مطابقة للمعايير المطلوبة، كما أن هناك أنواع مختلفة من أخطاء الترجمة. بالإضافة إلى ذلك فإن هناك اختلاف بسيط في جودة الترجمة وعدد الأخطاء لصالح ترجمة جوجل مقارنةً بترجمة مايكروسوفت، كما تبين أن نظام الترجمة الآلية العصبية لا يزال بحاجة إلى الكثير من التدريب ويحتاج إلى بناء المزيد من مجاميع النصوص العربية.

الكلمات المفتاحية: الترجمة الآلية، التقييم، الدقيق، الأخطاء، الإنجليزية-العربية، الملخصات، البحوث التربوية.

Introduction

The quality of machine translation is a very interesting field of research. Accompanying the great advancement in this field, there are many heated discussions regarding the quality of machine translation. In the literature, there are proposals that machine translation has achieved parity with professional human translation (Hassan et al., 2018; Barrault et al 2019). On the other hand, there are proposals that states that such parity has not been achieved (Läubli et al, 2018; Toral et al 2018; Freitag et al, 2021).

Regardless of the debates, there is no doubt that machine translation is advancing and that high-quality translations are performed by machine translation. However, it is also undeniable that there is still a gap between machine translation and professional human translation. Recent studies that performed error analysis have come out with a comparatively long list of errors (Popovic, 2021; Kocmi, 2022).

The topic of translation quality is of more interest when we discuss the translation between a pair like English and Arabic and when it involves a case of academic translation. English and Arabic languages show numerous morphological and syntactic variation. It should be indicated that Arabic is a language that has a very rich inflectional system which leads to NLP challenges that needs to be handled by using morphological analysis and tokenization tools for processing (Attia, 2007; Farghaly & Shaalan, 2009; Khalifa et al., 2016; Salloum & Habash, 2022, among others). Accordingly, it is expected that open MT systems face challenges, and the translation from a language with poor morphology to another with rich morphology is borne to be riddled with errors.

The case of translating academic writing is also interesting due to the nature of the language used which is naturally filled with specialized vocabulary. This poses a real challenge when a language like Arabic is involved. That is due to the scarcity of annotated Arabic corpora compared to other prominent languages. Most of the available Arabic corpora are primarily taken from media or are related to the political field. (For more details on this topic, check MeEntry et al. (2009).)

It should be indicated here that there is a growing interest in using machine translation for various purposes among academics, one of the most prominent of which is to translate research paper abstracts (as discussed in 2.2). However, due to the importance of abstracts since it is a summary of the whole paper, a poor-quality MT output is not of any help to the scholars and only a high-quality translation is acceptable in the case of abstracts.

This study, therefore, aims to investigate the quality of neural machine translation when translating research paper abstracts in education from English to Arabic. The study evaluates the translation quality of Google Translate and Microsoft Translator systems when translating research paper abstracts from English to Arabic. It also presents a classification of errors that occur when translating such abstracts from English to Arabic.

The study provides a fine-grained error analysis of NMT when translating research paper abstracts in the education field. It is without doubt that fine-grained analyses of MT errors contribute effectively to the development of MT, and accordingly, they contribute in the production of high-quality MT since they highlight the points of weakness and strength of MT systems. It will help in increasing the productivity of post-editors, as well as saving time and effort by providing an insight on the nature of the MT issues. The fine-grained analysis is performed on professional texts which is an aspect that the Arabic MT literature requires.

Literature Review

2.1 Machine Translation

Machine translation has received great interest recently and it has developed greatly in the few past years. After the emergence of the Neural Machine Translation (NMT) system, which is considered to be a great breakthrough in the field of MT, some research work has been done to evaluate it. The quality of translation produced by NMT systems is compared to the quality of translation provided by preceding systems, the most prominent of which is Phrase-Based Machine Translation (PBMT).

In the research that has been performed to compare the performance of both NMT and PBMT systems, it has been indicated that NMT outperforms PBMT in many aspects. An analysis of these system performance on English to German has been done by Bentivogli et al (2016) which has concluded that NMT minimizes editing effort and improves greatly in terms of inflection and word order. Other analyses have been performed on these systems by Toral and Sanchez-Cartagena (2017). These analyses have concluded that NMT surpasses PBMT in terms of inter-system variability, fluent outputs, and reordering. Klubicka et al (2017) have also performed an analysis on English-Czech which has showed that NMT is better in handling agreement and in producing fluent and grammatical language. However, it is pointed out that NMT degrades faster with sentence length as indicated in Bentivogli et al (2016). It is one of the NMT challenges that has been pointed out in Koehn and Knowles (2017) as well.

In general, recent research in the field has also indicated that NMT system outperformed PBMT and other Statistical Machine Translation (SMT) systems. Sennrich and Zhang (2019) have proposed that NMT system comes first in low-resource languages on generic domains. Ahmadnia and Dorr (2020) have also stated that NMT has surpassed SMT systems in low-source domains with specific data. Saunders (2022) has also indicated that NMT systems benefit from domain adaptation to achieve better performance with limited training data.

However, the topic of the quality of NMT is still controversial. 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 is close 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).

In regard to Arabic MT, Zakraoui et al (2021) have performed a survey on the challenges of Arabic MT. They have observed that research work in Arabic MT has been performed on both linguistic and technical issues with more focus on the linguistic ones. They also observed that NMT is always better than SMT and that research on Arabic NMT has increased recently. The survey has also shown that some efforts have been done to evaluate the effectiveness of MT.

It is true that some research has been done to evaluate the effectiveness of NMT or MT in general. However, those do not seem to be enough to make such an evaluation. In this regard, Ameur et al (2020) have 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 to English. Translating English to Arabic has been of secondary significance. This is really a big deal since it seems that more challenges 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. Ameur et al (2020) concluded that there are still a lot of Arabic-related linguistic problems that need a lot of investigation.

It can be stated here that despite the great development of MT and the superiority of NMT systems in comparison to their predecessors, it is still far from providing high quality translation. High-quality translation is more required when it comes to professional texts. Therefore, more investigation should be made in this area. Fine-grained error analyses are still needed in order to develop the MT systems. Arabic MT studies are still required in this aspect as well. Therefore, the study will be a great addition to the field of MT especially to error analysis of professional texts.

Therefore, the focus of the study is mainly on English to Arabic and where the translation quality is evaluated and errors are classified. The texts under investigation are of special nature which require the use of specialized language and terms. Therefore, a high-quality translation is required and the professional evaluation and the detailed analysis of errors performed in this study will undoubtedly provide a great insight to the development of MT.

2.2 Abstracts and MT

An abstract, as stated in Gastel & Day (2022), should be considered as a miniature version of the paper where a brief summary of the main sections of the paper is provided. This summary includes the introduction, methodology, results, and discussion. “A well-prepared abstract enables readers to identify the basic content of a document quickly and accurately, to determine its relevance to their interests, and thus to decide whether they need to read the document in its entirety” (American National Standards Institute, 1979, as cited in Gastel & Day, 2022. p. 59).

Since an abstract is a very important part of a research paper, it should adhere to the scientific writing norms and it should be clear, concise and readable. Accordingly, a translation of an abstract should also meet such standards. The translation provided should therefore be of high quality to ensure that the translated abstract is clear, concise and readable.

As discussed in Olohan (2016), translation of abstracts can be for both publishing and non-publishing purposes. In the case of publishing purposes, one can find many Arabic journals that publish essays in English but an Arabic abstract is also required. In the case of non-publishing purposes, researchers need to translate English abstracts in cases the researchers have no access to the English language. That becomes necessary when they need to consult a paper written in English considering the fact that many journals are published in English and scholars tend to use English to publish their work in the various fields of knowledge.

2.3 Translation Quality Assessment

Translation quality assessment (TQA) is a complex issue that has been debated by both academics and industry professionals. In academia, TQA is typically concerned with developing measures that can demonstrate a change in quality either by showing improvement in a translation compared to previous work or between different translation processes. However, in industry, the aim is to ensure that a specified level of quality is met. (Castilho et al, 2018)

There are many ways to assess the quality of translations (TQA), both in research and in industry by humans. The most common approach is to evaluate adequacy and fluency. Adequacy, also known as accuracy or fidelity, is a measure of how well a machine translation (MT) output conveys the meaning of the original source text. It is often used in conjunction with fluency, which is a measure of how well the MT output adheres to the rules and norms of the target language. In other words, adequacy is about whether the translation is accurate, while fluency is about whether the translation is natural and easy to read. According to Arnold et al. (1994), grammatical errors, mistranslations, and un-translated words can make it difficult to understand a text or speech, which they refer to as fluency. Reeder (2004) supports this view by finding that incorrect pronouns, inconsistent prepositions, and incorrect punctuation were all predictors of low fluency in experimental conditions. There are also other factors besides adequacy and fluency such as readability, comprehensibility, usability, and acceptability which can also be considered, especially for machine translation (MT) output.

Chatzikoumi (2020) classifies the methods of machine translation evaluation as follows. Automated evaluation uses machines to reach at MT outputs without any human involvement, while human evaluation involves humans in the evaluation process. Automated evaluation techniques can be divided into three types:

- Reference translation-based metrics: These metrics compare the MT output to a human translation of the same text, called a reference translation. The more similar the MT output is to the reference translation, the higher the score.
- Quality estimation (QE) metrics: These metrics classify the MT output into different quality levels. QE metrics are not evaluation metrics in themselves, but they are used as proxies for them.
- Diagnostic evaluation based on checkpoints: These metrics identify errors or weaknesses in the MT output.

Human evaluation techniques can also be divided into two categories:

- Directly expressed judgment (DEJ)-based: judges in this evaluation are required to assess the quality of the MT output. The judges have to make an assessment. DEJ-based techniques are more subjective than non-DEJ-based techniques.
- Non-DEJ-based: the process in this evaluation is task oriented (such as classifying errors or answering questions about the content of the text).

Despite the fact that human evaluation is more expensive, effort-intensive, and time-consuming, the researchers here prefer using human evaluation since it is more accurate and suits the purpose of the study. It is also the believe that automatic evaluation is just a substitution of human evaluation (Popović, 2020). The feedback provided by human evaluation is considered to be more accurate and comprehensive when compared to automatic evaluation (Chang et al., 2023). Two types of human evaluation are performed in this study.

Methodology and Results

3.1 Data

The research team selects 25 English research paper abstracts in education from well-known Scopus scientific journals published in America and Britain to ensure the quality of the source texts. The abstracts used in the study are selected from recent issues and are translated into Arabic using both Google Translate and Microsoft Translator. Five abstracts are used in the pilot study and 20 abstracts are used for the final evaluation and error annotation process.

3.2 Annotators / Evaluators

The evaluation and the annotation of errors are carried out by a team of four professional annotators who have a long experience in the field of translation and annotation. The team members are native speakers of the target language (Arabic) and have near native fluency of the source language (English). They are also experienced in the field of education. That is to ensure the integrity of the results.

3.3 Pilot Study

Prior to the analysis, a pilot study was performed. Five abstracts were provided to the team with evaluation and annotation guidelines. The performed evaluation and error annotations were thoroughly reviewed by the research team and feedback was provided to the team in case of any misunderstanding of the process or the guidelines. The research team also clarified any doubts and answered any questions raised by the annotators. It should be mentioned that the pilot study helps in identifying the types of errors that occur in the translation, which greatly helps in narrowing down the error span and facilitating the error annotation process.

3.4 Human Evaluation

It is worth mentioning that MT evaluation was largely devoted to sentence-level evaluation; however, the value of document level evaluation has been highlighted recently (Toral et al., 2018; Läubli et al., 2018; Läubli et al., 2020; Graham et al., 2020; Toral, 2020, among others). Therefore, two main suggested types of evaluation have been recommended; full document-level evaluation as presented in Läubli et al. (2018) and Läubli et al. (2020) and segment-level evaluation as proposed in Graham et al. (2019) and Graham et al. (2020).

On the other hand, pairwise comparison was suggested in Läubli et al. (2018) for evaluating both fluency and adequacy where each text pair is compared. The text that appropriately conveys the meaning will have higher accuracy and the text with better language will be the one with higher fluency. Some suggestions were also recommended in Läubli et al. (2020) to enhance the effectiveness of the evaluation. The framework introduced in Läubli et al. (2018) also supplied substantial evidence refuting the claim that MT is equivalent to human translation. Recommendations provided in Läubli et al. (2018) have been adopted in the large-scale evaluation campaign at WMT 2019 (Barrault et al., 2019). It is also indicated that Läubli et al. (2020) recommendations represent great progress in the evaluation field (Poibeau, 2022).

Another efficient method, regarding this aspect, which has been suggested in the reported literature, is segment-level evaluation. In this evaluation, a direct assessment is supplied for sampled segments (Graham et al., 2019; Graham et al., 2020). Nevertheless, it has been demonstrated that segment-level evaluation tends to downplay the disparities between human translation and MT (Barrault et al., 2019; Läubli et al., 2020). This serves as a compelling rationale for avoiding using this evaluation method in the present study.

In this study, however, since pairwise ranking is not the only intention and an evaluation of the translation quality provided by each system is examined, a professional evaluation is performed based on scalar quality metric (SQM) (Freitag et al., 2021). The study employs the SQM which uses a 0-6 scale as follows.

- **6: Perfect Meaning and Grammar:** The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.
- **4: Most Meaning Preserved and Few Grammar Mistakes:** The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.
- **2: Some Meaning Preserved:** The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.
- **0: Nonsense/ No meaning preserved:** Nearly all information is lost between the translation and source. Grammar is irrelevant.

The professional annotators were provided by the source texts (English abstracts) and their correspondent translations. Presented with the SQM guidelines, the professional annotators provided an evaluation for each abstract. Given the short nature of the abstracts, the annotators evaluated the translation outputs of whole texts and not just selected segments. It should be mentioned that this method is used in the WMT 2022 General Machine Translation Task (Kocmi et al, 2022) and WMT 2023 General Machine Translation Task (Kocmi et al, 2023). It is also used in the IWSLT 2022 human evaluation campaign (Anastasopoulos et al., 2022). It is proposed that the scores are stabilized when using these guidelines. It should be noted here that, as opposed to WMT, the evaluation here do not include giving a 1-100 score. The annotating team simply tick a score that fall between 0 and 6.

The evaluation here is performed by the team of four professional annotators as mentioned earlier. The evaluation results are shown in Table 1 below.

Table 1: Results of annotators' evaluation

MT System	Mean	Standard Deviation
Google Translate	4.23	0.66
Microsoft Translator	3.81	0.59

According to Table 1 above, Google Translate achieves an average score of 4.23 with a standard deviation of 0.66. On the other hand, the average score achieved by Microsoft Translator is 3.81 with a standard deviation of 0.59.

3.5 Error Analysis

The study performs error taxonomy. The taxonomy of the annotated errors in the study is guided by the one provided by Multidimensional Quality Metrics (MQM) introduced in Lommel et al. (2014). The typology of errors provided by MQM classified translation errors into eight dimensions: terminology, accuracy (adequacy), linguistic conventions (fluency), style, locale conventions, audience appropriateness, design and markup, and dimensions. Such dimensions are defined and classified further (https://themqm.org/the-mqm-full-typology/).

The errors detected fall under the following general dimensions: terminology, accuracy, linguistics conventions, style, and custom. These categories are classified further as it is indicated in the following section. Table 2 provides the number of errors of each category and subcategory. What follows is explanation and examples of the annotated errors in the abstracts under investigation.

3.5.1 Error Classification

Terminology: Errors occur when a term fails to adhere to the established standards of the subject field or organizational terminology, or when the target content contains a term that is not the accurate and normative equivalent of the corresponding term in the source content.

- **Inconsistent Use of Terminology:** This refers to the multiple terms which are used for the same concept where consistency is desirable. There are a number of cases of those errors which have been detected. For example, the word "loaded" in loaded moments, have been translated as "المشحونة" and as "المحملة" in another sentence of the same abstract.
- **Wrong Term**: This points to the use of a specific term which is not the term that a professional translator can use, or which can cause conceptual mismatch. There are numerous examples regarding this error. "Escalating and deescalating", for example have been translated as "التصعيد وخفض التصعيد" instead of, for illustrative purposes, "تأجيج وتهدئة". "Reading frequency" in the sentence "One often used approach to increase students' reading frequency is investing in independent silent reading (ISR) at schools" has been translated as "تكرار القراءة" as an alternative for "وتيرة القراءة", for the sake of argument.

Accuracy: Errors arise when the intended meaning of the target content deviates from the propositional content of the source text due to distortions, omissions, or additions to the message. Under this dimension the errors are classified further into the following:

- **Ambiguous Target Content**: This represents the case in which a specific term can be potentially interpreted in more than one way. Some cases have determined, for example, "separating the summer" have been translated as "فصل الصيف" which could be interpreted as "summer season" in the target text.
- **Ambiguous Source Content**: This relates to the source content which could be translated inappropriately in the target text. Few cases have been identified in the annotated abstracts. The term "state" have been misconstrued and interpreted as "الدولة/ الولاية" rather than "الحالة" as is required by the context. The term "scholarship", is another example that has been translated as "منحة"as a substitute for "بحث".
- **Overly Literal**: This pertains to the word for word equivalence in the target language when an idiomatic translation is required. In the translated abstracts, for example, the word "color" in "color-evasive and pathologizing discourses" has

been translated literally in this context as "الخطابات المراوغة للألوان والمرضية" in place of "الخطابات المرضية التي تتجنب الحديث عن العنصرية", for the sake of argument. The term "turn" in "take a descriptive turn" has been translated as "اتخاذ منعطف وصفي" as a different choice from "اتخاذ منحىً وصفياً.

- **Untranslated**: This points to a segment that was supposed to be translated but has been omitted in the translation. There are many examples which have been noted in the translated abstracts such as "ISR". It has been written as it is in English without being translated into Arabic Language.
- **Omission**: This pertains to not translating a content word in the translated abstract while such word has been present in the source text. There are certain instances of such error in the translated abstracts. For example, "(de)escalate" has been translated as "تهدأ" ignoring the brackets which indicate that there are two opposite words which should be translated to the target language.

Linguistic Conventions (Fluency): Errors which are related to the structure of the text including grammar and idiomatic expressions. Under this dimension the errors are classified further into the following:

- **Word Form**: This represents choosing the inappropriate morphological variant of a word, which include tense, agreement, and part of speech. There are numerous cases which have been spotted. For instance, the phrase "does not guarantee that students read" has been translated into " فمجرد تخصيص وقت للقراءة لا يضمن أن الطلاب القراءة". Instead of using "يقرأون", the "القراءة" was used. For more clarification, the phrase "the COVID-19 pandemic has caused" has been translated as "جائحة كوفيد-19 تسبب في" where "تسببت" should have been used.
- **Word Order**: This signifies the non-compliance of the word order of the translation to the norms of the target language. Several occurrences of such error have been determined in the translated abstracts. The phrase "rapid skill development" has been translated as "لتنمية المهارات السريعة". A more faithful translation would have been "التنمية السريعة للمهارات".
- **Incorrect Function Word**: This concerns the error of using incorrect function word, which is essential for showing relationships between content words and conveying clear meanings. Multiple cases of this error have been found in the translated abstracts. For example, "in 489 German university instructors" has been translated into "في 489 مدرساً جامعياً ألمانيًا ". A more accurate translation would be "لدى 489 مدرساً جامعياً ألمانيًا".

- **Missing Function Word**: This represents the case when a function word is required but it is not present in the target text. Many cases of this error have been found in the translated abstracts. The sentence " Significant attention and legislation have been directed to assessment intervention for students with word-level reading disability.", for example, has been translated as " تم توجيه اهتمام كبير وتشريعات كبيرة التدخل التقييمي." However, for a more correct translation, a preposition like "نحو" should have been used as follows " تم توجيه اهتمام كبير وتشريعات كبيرة نحو التدخل التقييمي ".
- **Extraneous Function Word**: This refers to using unnecessary function word in the translation. Some examples have been observed in the translated texts. The phrases "in the field related to (a) how learners are categorized, (b) what is being learned" has been translated into " في المجال المتعلق بـ (أ) كيفية تصنيف المتعلمين، (ب) ما هو التي يتم تعلمها،". For flawless rendition, the words "هو التي" should be removed. The sentence should be "في المجال المتعلق بـ (أ) كيفية تصنيف المتعلمين، (ب) ما يتم تعلمه ".
- **Punctuation**: This refers to the incorrect use of punctuation marks based on the target language rules. There are only limited cases which have been detected in the translations. For instance, the sentence "students are typically presented with new information through several modalities, such as language and images" has been translated into "يتم تقديم معلومات جديدة للطلاب عادةً من خلال عدة طرق، مثل اللغة والصور". However, there is no need for the coma according to the rules of Arabic Language in such a sentence.
- **Spelling**: This concerns the errors which are related to miswriting words. Only one case has been discovered in the translation. The phrase "linked to their well-being" has been rendered in the translated text as "مرتبطة برفاهتهم" in place of "رفاهيتهم".
- **Duplication**: This refers to using the same word, phrase, or sentence more than once though it is mentioned only once in the source text. Only one case has been identified in the translated abstracts. The source text "(a) 189 emergent bilingual students receiving services for English language development (ELD); (b) 374 reclassified bilingual students who had exited ELD programs;" has been rendered in the translation as follows "(أ) 189 طالبا ناشئا ثنائيي اللغة يتلقون خدمات لتطوير اللغة
؛ (ELD) ؛ (ب) 189 طالبا ناشئا ثنائيي اللغة يتلقون خدمات لتطوير اللغة الإنجليزية (ELD) الإنجليزية
؛ (ب) 189 طالبا ناشئا (ELD) (ب) 189 طالبا ناشئا ثنائيي اللغة يتلقون خدمات لتطوير اللغة الإنجليزية
؛ (ب) 189 طالبا ناشئا ثنائي اللغة يتلقون (ELD) ثنائي اللغة يتلقون خدمات لتطوير اللغة الإنجليزية
؛ (ب) 189 طالبا ناشئا ثنائي اللغة يتلقون خدمات لتطوير اللغة (ELD) خدمات لتطوير اللغة الإنجليزية

؛ (ELD) ؛ (ب) 189 طالبا ناشئا ثنائي اللغة يتلقون خدمات لتطوير اللغة الإنجليزية (ELD) الإنجليزية (ب) 189 طالبا (ب) إعادة تصنيف 374 طالبا ثنائيي اللغة كانوا قد تخرجوا من برامج التعليم في اللغة الإنجليزية؛". The item (a) is translated then the same item is repeated several times with a different item number, (ب).

- **Inconsistency in using tense**: This concerns of not using the same tense across the text. One case has been found in the translated texts. The source text is "We draw on data from 117 countries to describe cross-national patterns in higher education attendance rates, disaggregated by wealth quintile and country income group. We then calculate four different indicators to quantify the size of wealth-based inequality in higher education attendance and completion for each country." The translated text is " نحن نعتمد على بيانات من 117 دولة لوصف الأنماط العابرة للحدود الوطنية في معدلات الالتحاق بالتعليم العالي، مصنفة حسب شريحة الثروة ومجموعة دخل الدولة. ثم قمنا بعد ذلك بحساب أربعة مؤشرات مختلفة لتحديد حجم عدم المساواة على أساس الثروة في الالتحاق بالتعليم العالي وإتمامه في كل بلد". It could be noticed the shift from using the present tense in ""نحن نعتمد to the past tense in "ثم قمنا بعد ذلك".

Style: This refers to the text errors which are grammatically appropriate; however, they are inappropriate due to exhibiting inappropriate language style or deviating from organizational style guides. A considerable number of instances of this error have been uncovered in the translated texts. For example, the sentence "The results serve as a foundation for future studies on how country-level factors and policies exacerbate or reduce wealth-based inequalities." has been translated as " وتخدم النتائج كأساس للدراسات المستقبلية حول كيفية تفاقم أو تقليص العوامل والسياسات على مستوى الدولة لأوجه عدم المساواة القائمة على الثروة ". The sentence is grammatically and meaningfully acceptable; however, stylistically in Arabic language the phrase "... تُعد النتائج بمثابة أساس" is more commonly used compared to "..... وتخدم النتائج كأساس" Another example for such error is "the field has become more strongly international in its orientation" which has been translated into "أصبح هذا المجال دوليًا بقوة في توجهه ". A more acceptable translation in Arabic would be "ذا توجه دولي مُلفت".

Custom: This dimension is included in MQM to accommodate other errors that do not fall under the previous seven dimensions. In this study, the “breaking up long sentences” error is included in this dimension.

- **Breaking up long sentences**: This refers to an error when a MT system fails to process a long sentence. It seems that there is a specific number of words for the length of a sentence to be translated. Therefore, the system ends the sentence at that limit, puts a period, and starts a new sentence. There are a limited number of cases regarding this error. The sentence in the source text "Following feedback, we assessed study participants’ real-time (i.e., state level) epistemic emotions (surprise, curiosity, enjoyment, confusion, frustration, anxiety) and achievement emotions (anger, pride) produced by high-confidence errors (i.e., incorrect answers a person was confident in)." has been divided into two sentences as " بعد الحصول على ردود الفعل، قمنا بتقييم المشاعر المعرفية للمشاركين في الدراسة في الوقت الحقيقي (أي على مستوى الدولة) (المفاجأة والفضول والاستمتاع والارتباك والإحباط والقلق) ومشاعر الإنجاز (الغضب والفخر) الناتجة عن أخطاء الثقة العالية (أي غير صحيحة). إجابات كان الشخص واثقا منها". This goes with what has been indicated in Bentivogli et al (2016), as explained earlier, regarding the fast degradation of neural machine translation with sentence length.

Table 2 below presents the type of error along with the number of the annotated errors of the translated abstracts from English to Arabic by Google Translate and Microsoft Translator based on MQM taxonomy.

Table 2: Number of Errors in the translated abstracts in Google and Microsoft

Dimensions	Types of Errors	No. of Errors in Google	No. of Errors in Microsoft
Terminology	Wrong Term	105	119
	Inconsistent use of terminology	14	8
Total of Terminology Errors		**119**	**127**
Accuracy	Ambiguous target content	5	4
	Ambiguous source content	3	3
	Overly literal	14	15
	Untranslated	16	22
	Omission	5	6
Total of Accuracy errors		**43**	**50**
Linguistic conventions (Fluency)	Word Form	29	33
	Word Order	13	11
	Incorrect FW	18	17
	Missing FW	19	21
	Extraneous FW	9	5
	Punctuation	5	5
	Spelling	1	0
	Duplication	0	1
	Inconsistency in using Tense	1	0
Total of Linguistic conventions (Fluency)		**95**	**93**
Style	Style	17	18
Custom	Breaking up long sentences	4	0
Total		**278**	**288**

3.5.2 Error Distribution:

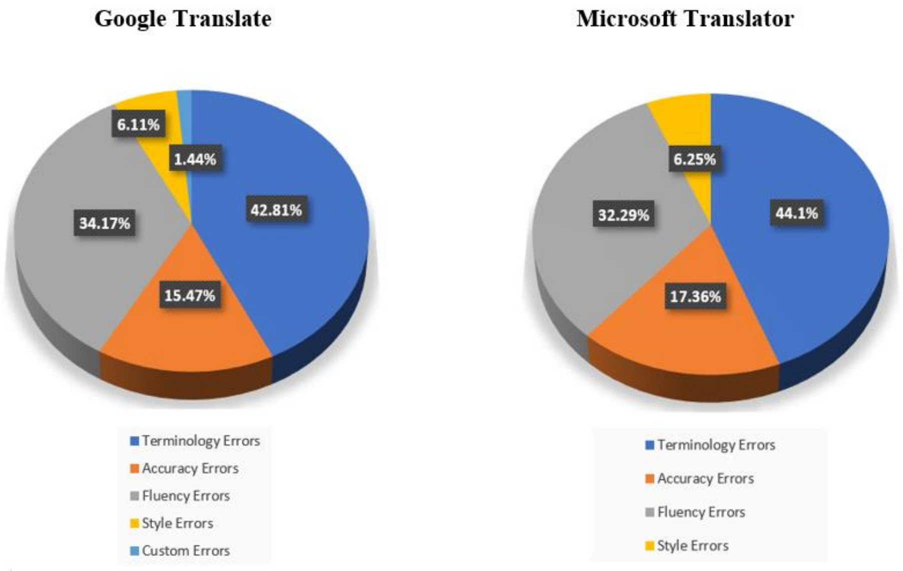
Upon examination of the aforementioned data, one can discern that the frequency of errors in Microsoft are higher than the frequency of errors in Google Translate. The annotated abstracts amount to 130 sentences, which indicates that the rate of error occurrence is 2.14 per sentence in Google Translate translated abstracts and 2.23 per sentence in Microsoft Translator translated abstracts.

The distribution of errors according to MQM main dimensions in Google Translate is as follows: 119 distribution errors (42.81% of the annotated errors), 43 accuracy errors (15.47% of the annotated errors), 95 fluency errors (34.17% of the annotated errors), 17 style errors (6.11% of the annotated errors), and 4 custom errors (1.44% of the annotated errors).

In Microsoft Translator translated abstracts, the distribution of errors are as follows: 127 distribution errors (44.1% of the annotated errors), 50 accuracy errors (17.36% of the annotated errors), 93 fluency errors (32.29% of the annotated errors), 18 style errors (6.25% of the annotated errors), and no custom errors.

The distribution of errors according to MQM main dimensions is represented in the figure below.

Figure 1 Error Distribution in Google Translate and Microsoft Translator



Discussion

This research paper investigates the quality of neural machine translation when translating research paper abstracts from English to Arabic by comparing the translations of Google Translate and Microsoft Translate. The evaluation of the translation quality showed that Google Translate achieved a slightly higher average score (4.23) compared to Microsoft Translator average score (3.81). This also comes in line with the distribution of errors where the error rate in the translation produced by Google Translate (4.23) is lower than the rate in the translation produced by Microsoft Translate (3.81).

However, it should be noted that neither systems meet the required standards. Both systems still generate translations that contain numerous errors. The high number of terminology and accuracy errors indicates that both systems are lacking enough training data for academic Arabic texts. On the other hand the high number of fluency errors indicates that both systems are still unable to capture all the structural divergences between English and Arabic.

Conclusion

The research findings revealed that there are many errors that occur when translating research paper abstracts from English to Arabic using Google Translate and Microsoft Translate. It is found that both systems have a higher number of fluency errors compared to accuracy errors, and they also exhibit a large number of terminology errors. Additionally, the exploration revealed that Google Translate slightly outperforms Microsoft Translate in terms of translation quality, achieving a higher average score and demonstrating fewer errors. Based on the explanations provided and considering the limited availability of Arabic corpora, it is evident that NMT system still requires extensive training, and the development of more Arabic corpora is necessary.

It should be noted here that the abstracts investigated are limited and they all taken from the education research domain. Therefore, most studies should be performed to cover other academic fields.

Acknowledgement

This research received grant no. (189/2022) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

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.

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.

Attia, M. (2007). Arabic tokenization system. *In Proceedings of the 2007 workshop on computational approaches to Semitic languages: Common issues and resources* (pp. 65-72).

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., ... & Zampieri, M. (2019). Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 1-61).

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).

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. *Translation quality assessment: From principles to practice*, 9-38.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, *26*(2), 137-16

El-Farahaty, H., Khallaf, N., & Alonayzan, A. (2023). Building the Leeds Monolingual and Parallel Legal Corpora of Arabic and English Countries' Constitutions: Methods, Challenges and Solutions. *Corpus Pragmatics*, *7*(2), 103-119.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, *8*(4), 1-22.

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.

Gastel, B., Day, R.A. (2022). *How to write and publish a scientific paper*, 9th edition. Greenwood, Santa Barbara, CA.

Graham, Y., Haddow, B., & Koehn, P. (2019). Translationese in Machine Translation Evaluation. *arXiv e-prints*, arXiv-1906.

Graham, Y., Haddow, B., & Koehn, P. (2020). Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 72-81).

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).

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.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., & Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, *67*, 653-672.

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791-4796). Association for Computational Linguistics.

McEnery, T., Hardie, A., & Younis, N. (2019). Introducing Arabic corpus linguistics. In T. McEnery, A. Hardie, & N. Younis (Eds.), *Arabic Corpus Linguistics*, (pp. 1–16). Edinburgh University Press.

Olohan, M. (2016). *Scientific and technical translation*. London: Routledge.

Poibeau, T. (2022). On “Human Parity” and “Super Human Performance” in Machine Translation Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6018-6023).

Popović, M. (2020). Informative Manual Evaluation of Machine Translation Output. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5059-5069).

Popović, M. (2021). On nature and causes of observed MT errors. In Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track) (pp. 163-175).

Sajjad, H., Abdelali, A., Durrani, N., & Dalvi, F. (2020). Arabench: Benchmarking dialectal Arabic English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5094-5107).

Salloum, W., & Habash, N. (2022). Unsupervised Arabic dialect segmentation for machine translation. *Natural Language Engineering, 28*(2), 223-248.

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

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).

Toral, A. (2020). Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. *arXiv preprint arXiv:2005.05738*.

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).

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).

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.