Journal of Namibian Studies, 34(2023): 2806–2822 ISSN: 2197-5523 (online)

Comparative Evaluation of Neural Machine Translation of fiction literature: A case study

Hanan Ibrahim[1], Linda Alkhawaja[2*]

Abstract

The development of machine translation has significantly improved the quality of translations. However, it is unfortunate that not all language pairings or genres benefit equally from this technology. This study investigates the quality of neural machine translation (NMT) output in the novel genre from English to Arabic languages. It examines two Machine Translation (MT) systems: Google Translate and Reverso, for translating quotes from Charles Dickens' novel, 'Hard Times.' The study aims to determine whether human translators can benefit from incorporating MT into their work and which MT system is valuable for translating this genre. To achieve this, a comparable corpus of 50 English quotes and their Arabic translations was used to assess the output quality of the two MT systems. The corpus was collected using CLiC (Corpus Linguistics in Context) software for literary analysis, and the evaluation was performed using the BLEU (Bilingual Evaluation Understudy) metric. BLEU compares MT outputs with professionally published human translations, generating scores for comparison. Based on the precision parameter used by BLEU, the results show that Google Translate slightly outperforms Reverso in producing high-quality output. These findings will help evaluate machine translation outputs in the novel genre compared to human translations. In conclusion, while the precision of human translators cannot be matched by the most advanced machine translation technology (NMT) in the novel genre, translators can still benefit from MT systems in their work.

Keywords: Neural Machine Translation, BLEU, Google Translate, Reverso, Translation Quality, Fiction Literature.

Introduction

Neural Machine Translation (NMT), introduced in 2014 and developed in 2017, has proven to be the most effective machine translation software thus far (Kenny, 2022). The new translation system exhibits a noteworthy improvement of 60% in reducing translation errors compared to its predecessor, Statistical Machine Translation (SMT), while also demonstrating a higher speed (Kenny, 2022). The described

[1] English Language Department, Al-Ahliyya Amman University, Amman, Jordan, h.ibrahem@ammanu.edu.jo

[2] English Language Department, Al-Ahliyya Amman University, Amman, Jordan, l.alkhawaja@ammanu.edu.jo

progress can be attributed to using an artificial neural network within the system. The network is purportedly designed after the human brain's neural structure, enabling the system to establish significant contextual associations among words and phrases. The system can establish these connections due to its proficiency in acquiring language rules, achieved by analysing numerous sentence examples from its database to detect recurring patterns. The machine uses the rules to generate statistical models, facilitating acquiring knowledge about sentence construction (Cullen, 2020).

The advancement of machine translation has significantly enhanced the quality of translations. However, not all language pairs benefit equally from this technology (Donaj and Kai, 2016, 2017). For instance, MT struggles with morphologically rich languages, particularly when translating from one morphologically simple language to another, such as from English to Arabic. Such language pairs are challenging to translate for MT and other language technology applications (Donaj and Kai, 2016, 2017).

Although neural systems demonstrate a high proficiency in translating specific text types, particularly those with formulaic structures and concise sentences, their capabilities remain limited (Rossi and Carré, 2022). This phenomenon can be attributed to the technical intricacies that underpin the respective systems. A substantial corpus of parallel sentences is necessary to train a system effectively. The system's performance will be optimised when trained on sentence types aligned with its intended translation tasks (Rossi and Carré, 2022).

The challenge faced by machine translation systems pertains to the fact that, in many instances, authors' styles in literature cannot be readily transferred, and no prior model exists upon which a system can be constructed (Thai et. al. 2022). One may believe that the significance of the authorial style is not paramount. However, when it comes to literature, form and function are bound together. According to Terry Eagleton, “there are certain obvious ways in which the idea of literature as self-expression is flawed, not least when it is taken too literally” (Eagleton, 2014: 136). The nature of literary language yields itself to different interpretations; this is true in the language of the novel genre and more so of the language of poetry which cannot be read the way a manual or a road sign is read.

This study attempts to assess the effectiveness of Google Translate and Reverso in translating mostly English quotes into Arabic. This study aims to further earlier research in the area of TS. In other words, several studies have already examined the output of various MT software in the English-Arabic language combination, including Google Translate, Systran, Babylon, The Translator, Sakhr, Al-Mutarjim TMAl-Arabey, and Systran (Hussein and Awab, 2016; Al mahasees, 2018; Jabak, 2019). Many studies (Popel et. al. 2020) tested and evaluated

the output of MT for specific domains such as media, news, politics, and business. Others (Al mahasees, 2018; Jabak, 2019; Zakraoui et al. 2021) studied MT techniques, problems, assessment, and analysis. Additional research (Belinkov and Durrani, 2017; Marouani et.al. 2018) has examined the shortcomings and faults of MT output, categorising them as, among others, lexical and syntactic issues and attributing them to the nature of the Arabic language which is highly complex. They discovered that there is still a need for betterment in Arabic due to several problems, including linguistic and syntactic errors. Nevertheless, none of the previous research looked at Reverso and Google Translate in terms of fiction.

Despite the MT's demonstrated ability to expedite the entire translation process, in this research, we are endeavouring to answer the following research questions. To what extent can human translators benefit from using MT systems in novel /fiction translation? Which one of them; Google Translate or Reverso is more effective to use in translating fiction? Is MT valuable for the translation of literary work? By understanding how MTs approach certain stylistic elements in literary texts, the computer-assisted study of literary texts and their translations may significantly contribute to the study of machine translation.

Theoretical framework

As the significance of MT is increasing as a mode of translation, assessing its quality becomes a critical consideration. The evaluation of translation quality is of two types: manual evaluation and automated evaluation (Rossi and Carré, 2022). As for the manual evaluation, it is predominantly conducted through human evaluation methodology. Translation and linguistics professionals evaluate machine translation (MT) output quality from two distinct perspectives. The initial perspective pertains to the level of adherence to the intended text and language norms, encompassing factors such as clarity and grammatical accuracy; a quality assessment aspect known as fluency. The initial text is not pertinent to the fluency assessment. The evaluators are only provided with the translation being evaluated, as the original data is not accessible to them (Rossi and Carré, 2022). Assessing fluency in a language requires a fluent expert in the target language. On the other hand, accuracy is the evaluation of how well the target text effectively communicates the source text's informative content. Evaluators can assess both the original text and the translated versions and frequently consider the sentence's context. The evaluators must master both languages to perform this task (Castilho, et.al. 2018). On a 5-point scale, adequacy and fluency are often scored. This study will not employ this

assessment because it is labor-intensive, costly, and fundamentally subjective.

We have seen significant advancements in automated MT assessment recently. The assessment of machine translation quality can be conducted through various means, depending on the assessment's objective and the available tools (Mauec and Donaj, 2019). Automated evaluation serves as a cost-effective substitute for human evaluation (Castilho, et.al. 2018). During the development of MT systems, they are commonly used to estimate the improvement of MT systems. These can also be utilised to compare various machine translation systems. It is important to understand the meaning of scores generated by automatic metrics when evaluating the quality of translations. They mainly rely on the idea that machine translation quality should inherently approximate human translation. The availability of human reference translation is a prerequisite for the use of automatic metrics. The evaluation of MT systems is conducted through a comparative analysis of the output against a reference translation. Evaluation metrics furnish assessment scores predicated on the reference translation that is most akin (Rossi and Carré, 2022).

Therefore, one feasible approach to assess the accuracy of a translation is to visually inspect the translated text and make a subjective determination as to its correctness (Castilho, et.al. 2018). In order to ensure dependable assessments, the evaluators must possess the necessary qualifications. As executed by proficient translators, manual evaluation is a costly and time-consuming process (Rossi and Carré, 2022). There is a requirement for automated metrics that are efficient and cost-effective while also providing a reliable estimation of human evaluations. Various effective metrics are utilised in the machine translation community, including but not limited to BLEU, NIST, METEOR, and TER. Reference translations are necessary for the computation of metrics as they facilitate the comparison of MT output with established translations, thereby generating scores for comparison. In cases where reference translations are accessible, the metrics above can expeditiously assess multiple systems' output, obviating the necessity for human involvement (Castilho, et.al. 2018). Despite numerous evaluation metrics for MT, BLEU (Bilingual Evaluation Understudy) remains the predominant measure of translation quality utilised by MT system developers (Maučec and Donaj, 2019). According to Warner (2022), this metric is most commonly used metric. Consequently, the BLEU metrics have been used for this research. BLEU scores evaluate the accuracy of translations by comparing MT translations with a human translation and then producing scores from 0 to 1, 0 to 10, or 0 to 100 with the higher number representing a better translation. In other words, a 100

score means that MT output and the human translation understudy are identical (Aiken, 2019).

A pertinent aspect to consider is that current MT systems operate at the level of individual sentences. This implies that they translate each sentence in isolation and discard it once they proceed to the next one (Kenny, 2022). Typically, this is a minor concern in the context of technical literature. In literature, the ability to recall ideas, metaphors, allusions, and images from earlier text sections is a crucial skill for a translator. While machines have made significant strides in this area, they still have a considerable distance to cover before they can match the proficiency of a human literary translator, as noted by Hadley (2020).

Even so, scholars are currently exploring MT's potential applications in the different literary genres. A recent investigation by researchers affiliated with the University of Massachusetts at Amherst sought to elucidate the reasons behind the comparative inadequacy of MT concerning human literary translations (Thai et. al. 2022). The researchers have compiled a dataset named PAR3, which comprises a minimum of two human translations for each source paragraph. In order to evaluate the effectiveness of MT in the realm of literature, the scholars utilised Google Translate to generate English renditions of the original paragraphs. These translations were then juxtaposed with human translations and presented to two distinct cohorts: proficient literary translators and English-speaking writers who only know one language. Notably, both cohorts exhibited a strong preference for human translations, as evidenced by the fact that human raters favored human translations over machine-translated versions in 84% of cases. The raters provided valuable insights that could enhance the capacity of machine translation for literary purposes (Thai et. al. 2022).

The researchers (Thai et. al. 2022) have identified five potential areas of improvement for machine translation based on the feedback received. Approximately 50% of the machine translation errors were attributed to a tendency towards overly literal text translation. Although these occurrences may not have constituted explicit errors, they frequently impeded the paragraph's coherence, resulting in a cumbersome reading experience. Furthermore, the absence of contextual information resulted in approximately 20% of the reported issues within the machine-translated paragraphs. The occurrence of translation errors can be attributed to various factors such as inadequate selection of words, imprecise or exact language, and what is referred to as "catastrophic" errors that render the translation completely invalid, such as misgendering a character. The raters utilised the insights above to devise an automatic post-editing model based on GPT-3, which was employed to modify the output generated by machine translation. The raters deemed the post-edited versions

more favorable than the unedited versions generated by Google Translate.

Google Translate is generally pretty accurate; it was founded in 2006 and has since grown to be one of the best MT tools, handling 133 languages now and 24 more in 2022 (Harby, 2023). Depending on the language pair and the type of material, accuracy varies, although some reports indicate that Google Translate can achieve 94% accuracy (Castilho e. al. 2019). Google's switch to NMT in 2016 marked a turning point regarding output quality. According to the tech behemoth (Harby, 2023), GNMT decreased translation errors by more than 60% for major language pairs. With "zero-shot translation," it was also no longer necessary to translate indirectly.

Google Translate fared well for European languages but less well for Asian languages, according to a 2011 accuracy assessment of 51 languages. Of course, that research is now out of date. According to a 2019 re-evaluation using the exact text and statistics, there had been a 34% improvement (Harby, 2023). Regarding reliability and accuracy, Google Translate is among the finest, especially for languages with scarce resources. According to Harby (2023), The MT evaluation program Intento placed Google Translate first among 18 other MT engines for practically all language pairs in 2022.

The Reverso system is also based on neural machine translation NMT. It is claimed (Reverso, 2023) that it can create texts that are simple to read even when your source material is intricate. It uses technologies like an artificial context dictionary to show you instances where the translated or original word was used in a real document so that you may better understand the language used in the phrase. This removes any uncertainty that could exist about the translation. Although many studies have examined Reverso's effectiveness for various language pairs, none have examined the system's performance for the Arabic-English pair.

Literature review

The current state of research on literary machine translation is limited, leaving uncertainties about the performance of modern machine translation systems, as noted by Warner (2022). Therefore, it is imperative to consistently scrutinise MT in the context of literary translation. Various research studies have been conducted to assess the effectiveness of machine translation applications and websites in translating academic content such as narratives, poetry, and dramatic works (Constantine, 2019; Abdulaal, 2022). According to Huang and Knight's (2019) findings, machine translation technologies can potentially be advantageous in translating Spanish literature into

English, despite minor mistakes and inaccuracies that Spanish software developers can readily address.

According to Chaeruman (2019) findings, there was a significant overlap in the quality of sentences produced by accredited professional translators and MT technologies, with 32.6% being nearly identical. Abdi and Cavus (2019) conducted an evaluation of different MT applications to translate Danish prose and poetry. The authors highlighted the potential utilisation of machine translation tools in literary translation. The authors argued that machine translation holds promising potential for language users concerning literary interpretation. Koehn (2020) conducted a research study to investigate the usability of machine translation in translating short stories from French to English. The author concluded that various lexical, grammatical, and structural errors negatively affected the translation quality. Absolon (2019) pointed out that insufficient attention has been paid to specific social and cultural inter-textual references. Therefore, Absolon concluded that machine translation technologies utilised for literary works are not that credible and reliable.

The present study analyses the utilisation of MT in literary works. MT is a potent instrument that can render literary works between languages. The utilisation of MT to translate literary texts at the document level, specifically those containing parallel paragraphs from world literature, was explored by Thai et al. (2022). The authors investigated the advantages of utilising MT for translating document-level texts, including heightened accuracy and efficiency. The authors also observed that MT can generate parallel corpora comprising translated documents in various languages. The efficacy of MT in translating English-Arabic texts was investigated by Beseiso et al. (2022). The researchers employed a methodology for evaluating translations based on semantics to compare the caliber of translations generated by machine translation systems and human translators. The study's findings indicated that MT can generate translations that are comparable in precision to those created by human translators. The authors concluded that MT could efficiently accomplish English-Arabic translation tasks.

Moreover, Abdulaal (2022) analysed the errors present in both machine and human translations of literary texts. He utilised corpus-based analysis to compare the errors committed by machine translation and human translators. The study's findings indicated that errors in machine translation were predominantly associated with lexical and syntactic factors. Conversely, instances of human error were predominantly attributed to a deficiency in comprehending the source material. Abdulaal concluded that human translators must

exercise caution and be cognizant of the possible errors that may arise during the translation of literary texts using machine translation.

Hadla et al. (2014) conducted an evaluation of the precision of MT in the translation of texts from Arabic to English. The researchers employed a comparative evaluation framework to assess the precision of MT and human translations. The study's findings indicated that MT achieved an accuracy rate of 79.3%, a performance level similar to that of human translations. The study's authors concluded that MT has the potential to be effectively utilised for Arabic-English translation assignments.

In summary, the present studies demonstrated MT's efficacy in literary translation works. MT can enhance the precision and effectiveness of translations at the document level while generating parallel corpora of translated documents in various languages. In addition, MT can generate translations comparable in precision to those generated by human translators, albeit with the possibility of encountering errors in lexical and syntactic matters. Based on the above studies, MT can potentially serve as a valuable resource for translating literary works. However, different aspects in different settings can influence the translation output and get different translation results. Outputs results are based on complex computations of training data which differ among MT systems, language pairs and genres or domains (Kenny, 2022). Consequently, this study is expected to reveal different results.

Methodology

Google Translate and Reverso Machine Translation were used to evaluate the machine translation of literature from English into Arabic. In this study, we used a sub-corpus from a large corpus that is available in English language on CLiC (Corpus Linguistics in Context) software. The focus was on finding quotes in dialogues in a novel written by Charles Dickens, namely Hard Times. Dialogue is essential in Dickens' novels because it reveals a character's traits, attitudes, emotions, and actions. CLiC software is user-friendly as it allows passages to be easily marked up for searches. The distinctive way CLiC searches speech in narrative fiction distinguishes it from other corpus tools, not just because of its technological advancements.

To create a collection of precisely annotated examples for our computation, we selected 50 random quotations from our text. The data was manually gathered from a published Arabic translation of the novel contrasted with our automated annotation in the original English quotes. Following that, MT systems were used to translate the source text segments. Finally, the MTs, human translations, and the

source text segments were submitted to the BLEU metric for review and analysis.

The data set must be carefully chosen as different types of data (e.g., genres, languages, styles, etc.) can lead to different results. In this instance, several data type subsets make up the evaluation set. In order to obtain more precise findings, we attempted to reduce the number of variables in our research by focusing on certain segments of a particular genre to derive more accurate results.

Data analysis and interpretation

The BLEU metric evaluates translation quality based on two distinct aspects: adequacy and fluency. It accomplishes this by measuring lexical precision through the calculation of word-level matches. Accuracy metrics such as BLEU-n, F-measure, Recall, and Precision, are utilised to assess translation quality with higher values indicating superior translation quality. To exemplify, the fundamental element of BLEU is the precision of n-grams. The evaluation metric measures the proximity between the output generated by the MT system and a professional human translation of the same text. The primary metric utilised by BLEU to differentiate between satisfactory and unsatisfactory MT outputs is the modified n-gram precision. This metric is determined by dividing the number of n-grams that correspond between the source and translated text by the overall count of n-grams present in the translated text that was assessed. The precision calculation is conducted individually for every order of n-gram, and subsequently, the precisions are aggregated through geometric averaging. The most frequently used definition for the maximum n-gram order is four, which refers to a sequence of four words. The BLEU metric calculates a modified precision score adjusted by a brevity penalty, also known as a length-based penalty. This penalty is applied to sentences shorter than the reference, discouraging their use, and the final scores exhibit a range from 0 to 1. Formulas 1 and 2 illustrate the operational mechanics of the calculation system utilised in the BLEU metric.

$$Bp = \begin{bmatrix} 1 & if\ c > r \\ \vdots & \vdots \\ e \quad 1 - r/c \cdots & if\ c \leq r \end{bmatrix}$$

Formula (2) demonstrates how the BLEU score is calculated from the BP stated in Formula (1).

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n \ \ log\, P_n\right)$$

Where N = 4 and uniform weights wn = (1/N) [2]

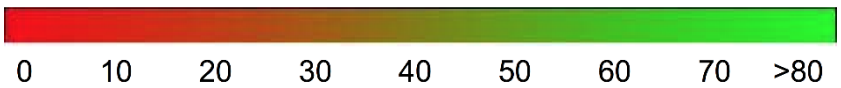
A cursory examination of the number of sentences exhibiting errors versus those not providing insight into the general calibre of the machine translation was done. Based on the cumulative score for the 50 segments in our dataset, the n grams for Google Translate; 1-gram 25.48, 2-gram 15.58, 3-gram 10.28 and 4-gram 6.18 are better than the n grams for Reverso; 1-gram 25.07, 2-gram 15.60, 3-gram 10.07 and 4-gram 6.65. Thus, the outcome of Google Translate exhibits a better value than the percentage ascertained from Reverso's output for identical segments. This implies that many of the examined segments contain errors and would require post-editing.

Tables 1 displays the results of an automated calculation of the accuracy of Google and Reverso machine translation systems for each of the four-gram sizes. By first computing the BP, selecting the helpful reference (i.e., the reference with more common n-grams), then computing the length, which is denoted by r (as shown in formula 1), and finally computing the total length of the MT translation denoted by c, the system combined the precision values in a single overall score (called BLEU-score).

Table 1: BLEU Score for Google MT and Reverso MT

BLEU Score:	**Google MT 6.18**				**Reverso MT 4.65**			
Precision x brevity:	6.18 x 100.00				4.65 x 100.00			
	1-gram	**2-gram**	**3-gram**	**4-gram**	**1-gram**	**2-gram**	**3-gram**	**4-gram**
Individual	25.48	9.52	4.48	1.35	25.07	9.71	4.20	0.46
Cumulative	25.48	15.58	10.28	6.18	25.07	15.60	10.07	4.65

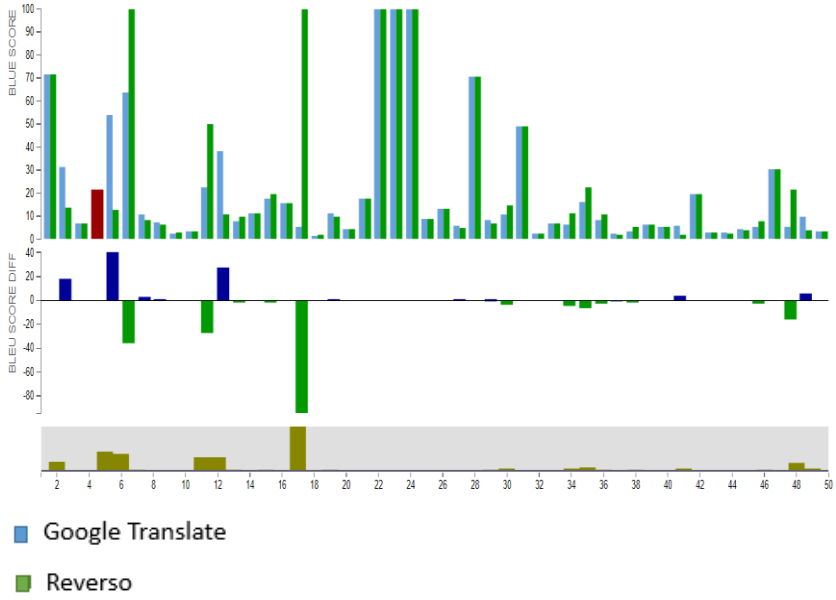
BLEU Score	**Interpretation**
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human



Interpretation of the BLEU score (Evaluating models | AutoML Translation Documentation | Google Cloud)

The bar chart illustrates BLEU score metrics for Google MT and Reverso MT. The scores range from anywhere on the scale between 0 - 100. The overall BLEU score metrics are 6.18 for Google Translate and 4.65 for. This means that the evaluation of Google's Translate output is higher than Reverso's output.

Chart 1: BLEU score metrics for Google MT and Reverso MT



It is important to remember that MT systems frequently commit serious translation mistakes, mainly when applied to complex genres like fiction. Although this assessment was based on a small corpus, the findings show that MT systems still require improvement. Tables 3 and 4 below present some random examples that are taken from the dataset.

Table 2: Examples from Google MT and their BLEU scores

Source Text	Human Translation	Google MT	BLEU Score/100	Length/ 1.00
Don't call yourself Sissy	لا تُطلقي على نفسك اسم سيسي	لا تطلق على نفسك اسم سيسي	53.73	1.00
Then he has no business to do it	لا يحق له أن يسميك هكذا	ثم ليس لديه عمل للقيام بذلك	2.38	1.29
my grandmother was the wickedest and the worst old woman that ever lived.	كانت جدتي أسوأ امرأة في هذه الدنيا	كانت جدتي أشر وأسوأ امرأة عجوز على الإطلاق	11.34	1.13

drink her four-teen glasses of liquor	أنها تشرب في سريرها أربعة عشر كأس م ن الخمر	تشرب كأسها من الخمور في سن ا لمراهقة	5.87	4.87
I were married on Eas'r Monday nineteen years in, long and dree	تزوجت منذ تسع عشرة سنة	تزوجت يوم الإثنين عشر سنة، طويلة ودرية	5.67	1.80
I coom home desp'rate	جئت إلى البيت وقد استبد بي اليأس	أنا أتعامل مع المنزل	2.94	0.50

Table 3: Examples from Reverso MT and their BLEU scores

Source Text	Human Translation	Reverso MT	BLEU Score/100	Length/ 1.00
Don't call yourself Sissy	لا تُطلقي على نفسك اسم سيسي	لا تسمي نفسك سيس	12.75	0.67
Then he has no business to do it	لا يحق له أن يسميك هكذا	ثم ليس لديه عمل للقيام بذلك	2.76	1.14
my grandmother was the wickedest and the worst old woman that ever lived.	كانت جدتي أسوأ امرأة في هذه الدنيا	كانت جدتي أخطر وأسوأ امرأة ع جوز عاشت على الإطلاق	9.98	1.25
drink her four-teen glasses of liquor	أنها تشرب في سريرها أربعة عشر كأس م ن الخمر	شرب أكواب الخمور الخاصة بها في سن المراهقة	4.87	0.89
I were married on Eas'r Monday nineteen years in, long and dree	تزوجت منذ تسع عشرة سنة	كنت متزوجة في يوم الإثنين تسع ة عشر عامًا، طويلة ورائعة	1.87	2.20
I coom home desp'rate	جئت إلى البيت وقد استبد بي اليأس	desp ' rate أقوم بتصنيف	2.91	0.75

As shown in table 2 and 3, the BLEU scores for both MT systems highlight serious errors in translation. Google Translate and Reverso both perform at similar levels, with Google Translate system showing marginally better scores on literary content. However, these scores do not mean that the output of both MT systems are of good quality especially if compared to a human translation.

After analysing MTs translation output, we can highlight some frequent errors committed by both systems. First, both systems produced severe meaning errors. For example, the sentence "drink her four-teen glasses of liquor" is translated by Google translate as "تشرب كأسها من الخمور في سن المراهقة" (Back translation: She drinks liquor during the teenage years". The same sentence is translated by Reverso as "شرب أكواب الخمور الخاصة بها في سن المراهقة" (Back translation: Drinking glasses of her own liquor in her teenage years".

Second, both systems mistranslated sentences that have a meaning in a given context. Although the translated sentences are readable, the reader cannot easily recover the original meaning without reading the

source sentence. For example, the sentence “Then he has no business to do it” is translated by both systems as “ثم ليس لديه عمل للقيام بذلك” (Back translation: Then he does not have a job/business to do this). However, this sentence in this particular context means that “he does not have the right to call you with this name”.

Third, both systems failed to translate some ambiguous words or phrases in the English source text that are unknown or rare words. For example, the sentence “I were married on Eas'r Monday nineteen years in, long and dree”. This sentence is translated by both systems as “تزوجت منذ تسع عشرة سنة” (Back translation: I married 19 years ago). The translations of both systems convey the original meaning of the source text but with a slight deviation.

Based on the analysis, approximately 60% of Google Translate output and 40% of Reverso output were found to be devoid of any errors, thus, would not require any further processing by a professional translator or post-editor. On that basis, one can conclude that human translators can benefit from using MT systems in novel /fiction translation but to a certain extent.

Translating literature is not only inconceivable for professional translators but also for the MT systems. That is mainly because considering discourse aspects in literary translation significantly contributes to the cohesion of literary texts. Also, MT systems are still not adapted to such aspects in literary texts. Therefore, we can say that MT is not that valuable for the translation of literary work. Despite the MT’s demonstrated ability to expedite the entire translation process, utilising a human translator remains necessary.

Implications for future research

The quality of machine translation is consistently enhancing. Notwithstanding this fact, several flaws exist in machine-translation output. In order to ensure translation quality, it is suggested that post-editing of machine translation output be incorporated into translation workflows. To bridge the gap in MT NMT research, it is imperative to undertake cross-disciplinary investigations incorporating the expanding corpus of pertinent knowledge in translation studies into the literature field's research. Cross-disciplinary research has the potential to enhance the quality of machine translation. Ultimately, the precision of human translators cannot be equalled by even the most advanced machine translation technology when it comes to the fiction genre. It is necessary to acknowledge that human translators constitute a significant component in the evolution of machine translation, serving not only as post-editors but also as instructors for MT systems within the educational domain. We highlight these issues as they merit further attention from researchers and policymakers.

Limitation of study

Acknowledging the study's limitation is crucial, which involves the use of a small corpus. This is mainly due to the limited availability of publicly accessible corpora in Arabic (Ahmed, 2022). A limited corpus can be advantageous if employed on a particular subset of the language or a diminutive, non-essential specimen of the language (Nesi, 2013). As in the case of this study, the small corpus comprises dialogue quotes from novels in a specific genre of the language.

Conclusion

Given the growing significance of artificial intelligence in contemporary society, adopting it as a translation mechanism is imperative to advance the field. The field of machine translation has made significant progress since its inception and is poised to assist human translators while alleviating the workload of human translators. NMT presents several prospects for translating challenging genres, including fiction. Upon achieving a better understanding of literary genres, it becomes feasible to undertake a comprehensive analysis of machine-translated literary texts compared to those translated by humans, as demonstrated in this study. This approach identifies additional challenges that machine translation may encounter at the textual level concerning literary texts. This is specifically important upon the recognition that tension exists "between form and content," which results in "discrepancy as part of the meaning of the work" (Eagleton, 2014, 3). Literary language is typically not so much practical or denotative as much as it is connotative.

The findings presented in this study indicate the potential significance and indispensability of conducting such analyses to advance toward an authentic literary machine translation.

The literature review indicates that studies concerning the utilisation of machine translation in the realm of fictional literature frequently overlook the intricacies of language and translation, which this study necessitates. The review's second point emphasises the need for increased awareness regarding the the distinct capabilities and constraints of machine translation in the context of fictional genres

Bibliography

Abdulaal, M. 2022. Tracing machine and human translation errors in some literary texts with some implications for EFL translators. Journal of Language and Linguistic Studies, 18(S1), pp.176-191.

Ahmed, A., Nashva A., Alzubaidi, A., Zaghouani, W., Abd-alrazaq, A., Househ, A. 2022. Freely Available Arabic Corpora: A Scoping Review. Computer Methods and Programs in Biomedicine Update. 2(2022), 100049.

Aiken, M. 2019. Paper An Updated Evaluation of Google Translate. Studies in Linguistics and Literature. Vol. 3, No. 3, p. 253-260.

Al-khresheh, S. and A. Almaaytah. 2018. "English proverbs into Arabic through machine translation", Int. J. Appl. Linguistics English Literature, vol. 7, no. 5, pp. 158.

Almahasees, Z. 2018. "Assessment of Google and Microsoft Bing translation of journalistic texts", Int. J. Lang. Literature Linguistics, vol. 4, no. 3, pp. 231-235.

AlMahasees, Z. 2020. "Diachronic evaluation of Google translate Microsoft translator and Sakhr in English-Arabic translation".

Al-Rukban, A. and A Saudagar, 2017. "Evaluation of English to Arabic machine translation systems using BLEU and GTM", Proc. 9th Int. Conf. Educ. Technol. Comput., pp. 228-232.

Belinkov, Y., N. Durrani, F. Dalvi, H. Sajjad and J. Glass. 2017. "What do neural machine translation models learn about morphology?", arXiv:1704.03471.

Beseiso, M., Tripathi, S., Al-Shboul, B. and Aljadid, R., 2022. Semantics based English-Arabic machine translation evaluation. Indonesian Journal of Electrical Engineering and Computer Science, 27(1), pp.189-197.

Castilho, et.al. 2018. Approaches to human and machine translation: Translation quality assessment. In: Translation Quality Assessment: From principles to practice. Edited by: Joss Moorkens, Sheila Castilho, Federico Gaspari and Steven Doherty. Springer: Switzerland.

Castilho, S., Gaspari, F., Moorkens, J. et al. 2019. Editors' foreword to the special issue on human factors in neural machine translation. Machine Translation 33, 1–7 (2019). https://doi.org/10.1007/s10590-019-09231-y

Constantine, P. 2019. Google Translate Gets Voltaire: Literary Translation and the Age of Artificial Intelligence. Contemporary French and Francophone Studies, Routledge. 23 (4). doi: 10.1080/17409292.2019.1694798

Cullen, A. 2020. Understanding Neural Machine Translation: How Artificial Intelligence "works" in Literary Translation. Available at: Literature and Translation | Goethe-Institut UK

Donaj G, Kačič Z. 2016. Language Modeling for Automatic Speech Recognition of Inflective Languages: An Applications-Oriented Approach Using Lexical Data. Springer, (5).

Donaj G, Kačič Z. 2017. Context-dependent factored language models. EURASIP Journal on Audio, Speech, and Music Processing, (1):6.

Eagleton, T. 2014. How to Read Literature. Yale University Press.

Hadley, J. 2020. Literary machine translation: Are the computers coming for our jobs?. Counterpoint, No.4. Counterpoint_2020_04_article_04.pdf (ceatl.eu)

Harby, A. 2023. How Accurate is Google Translate? Available at: How Accurate is Google Translate? - Slator

Hussein, S. and S. Awab, 2016. "Evaluation of Google and Bing online translations of verb-noun collocations from English into Arabic", J. Mod. Lang., vol. 25, pp. 35-59.

Jabak, O. 2019. "Assessment of Arabic-English translation produced by Google translate", Int. J. Linguistics Literature Transl., vol. 2, no. 4, pp. 10.

Johnson, M., Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, et al. 2017. "Google's multilingual neural machine translation system: Enabling zero-shot translation", Trans. Assoc. Comput. Linguistics, vol. 5, pp. 339-351.

Kenny, D. 2022. Human and machine translation. In: Machine translation for everyone: empowering users in the age of artificial intelligence, edited by Dorthey Kenny. Language science press: Germany.

Madi, N. and H. Al-Khalifa. 2020. "Error detection for Arabic text using neural sequence labeling", Appl. Sci., vol. 10, no. 15, pp. 5279.

Marouani, M. T. Boudaa and N. Enneya, 2018. "Incorporation of linguistic features in machine translation evaluation of Arabic", Proc. BDCA, pp. 500-511.

Maučec, M. and Donaj, G. 2019. Machine Translation and the Evaluation of Its Quality. In: Natural Language Processing - New Approaches and Recent Applications. DOI: 10.5772/intechopen.89063

Popel, M., Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, et al. 2020. "Transforming machine translation: A deep learning system reaches news translation quality comparable to human professionals", Nature Commun., vol. 11, no. 1.

Reverso. 2023. Reverso Corporate | Translation technologies, NMT, AI

Rossi, C and Carré A. 2022. How to choose a suitable neural machine translation solution. In: Machine translation for everyone: empowering users in the age of artificial intelligence, edited by Dorthey Kenny. Language science press: Germany.

Thai, K., Karpinska, M., Krishna, K., Ray, B., Inghilleri, M., Wieting, J. and Iyyer, M. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. arXiv preprint arXiv:2210.14250.

Warner, A. 2022. Humans still beat machines when it comes to literary translation. Available at: MultiLingual

Zakraoui, J., M. Saleh, S. Al-Maadeed and J. M. Alja'am, 2021. "Arabic Machine Translation: A Survey with Challenges and Future Directions," in IEEE Access, vol. 9, pp. 161445-161468, doi: 10.1109/ACCESS.2021.3132488.