ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 14, No. 2, pp. 347-357, February 2024
DOI: https://doi.org/10.17507/tpls.1402.05

Unveiling the New Frontier: ChatGPT-3 Powered Translation for Arabic-English Language Pairs

Linda Alkhawaja
English Language Department, Al-Ahliyya Amman University, Amman, Jordan

Abstract—**This study evaluates the aptitude of ChatGPT for Arabic-English machine translation. The main objective of this research is to scrutinize the quality of ChatGPT's translations and compare its performance against machine translation systems, such as Google Translate, which are intricately tailored for translation purposes. In addition, the study seeks to investigate the potential integration of ChatGPT into translation workflows. Furthermore, we aspire to discern whether ChatGPT's translation efficacy harmonizes with or diverges from the profound finesse exhibited by human translation expertise. To accomplish this, a comparable corpus of 1000 English sentences and their corresponding Arabic translations was employed to evaluate the translation outputs of both machine translation systems alongside a human translation reference. The corpus was sourced from Tatoeba, an open online platform and underwent electronic assessment using the BLEU (Bilingual Evaluation Understudy) metric. The results indicate a marginal advantage of ChatGPT over Google Translate in delivering high-quality translations. Upon evaluating the corpus, we ascertain that ChatGPT performs impressively well compared to specialized translation systems like Google Translate. However, despite these promising findings, it is essential to acknowledge that even the most advanced machine translation technology, ChatGPT, cannot currently match the proficiency of human translation, at least not in the near future.**

Index Terms—**ChatGPT-3-powered translation, machine translation, Google Translate, comparative study**

I. Introduction

Machine Translation (MT) stands as a critical field of exploration within the realm of natural language processing, and it has garnered significant attention in recent times. The core objective of MT involves the automated conversion of textual information from one language to another, facilitated by computational means. Hence, a proficient translation system must possess robust capabilities in comprehending and generating language, ensuring the production of apt and fluid translations. Previous studies (Liu et al., 2019; Guo et al., 2020) reveal that language models have the potential to improve source text comprehension through translation systems, but they face difficulties in producing perfect translation. Notably, ChatGPT has proven to be exceptionally skilled in comprehending and producing natural language, as seen by its ability to do so in a variety of contexts and produce responses that are human-like. As a result, the examination into using ChatGPT in the field of translations appears to be an intriguing and promising direction for further research.

Automatically translating text from one language to another is the focus of neural machine translation (NMT), a key task in the field of natural language processing (NLP) (Kocmi & Federmann, 2023; He et al., 2022; Stahlberg, 2020). Despite extensive research over the years, machine translation still grapples with several hurdles. These include tackling idiomatic expressions, managing translations for languages with limited resources, addressing uncommon words and upholding the flow and coherence of the translated text (He et al., 2022). The recent emergence of Large Language Models (LLMs) like ChatGPT (Wei et al., 2022; Ouyang et al., 2022; Chen et al., 2021; Brown et al., 2020) has significantly propelled the advancements in machine translation. Notably, LLMs can accomplish zero-shot machine translation, where a model is capable of translating between language pairs it has never been explicitly trained on, at a level comparable to robust, fully supervised MT systems. Moreover, these LLMs find utility across various applications beyond machine translation (Jiao et al., 2023; Wei et al., 2022; Wang et al., 2023).

ChatGPT, an AI chatbot developed by OpenAI and introduced in November 2022 (Ouyang et al., 2022), can comprehend instructions within prompts and furnish comprehensive replies. As per the official website (OpenAI, 2023), ChatGPT can respond and engage in subsequent queries, acknowledge errors, contest flawed premises and rebuff inappropriate solicitations within the conversational framework. The platform amalgamates diverse proficiencies in natural language processing, encompassing areas like addressing queries, weaving narratives, employing logical reasoning, debugging code, machine translation, and more. We are particularly focused on evaluating ChatGPT's performance in machine translation tasks, specifically comparing its performance with that of Google Translate.

We have formulated a set of research questions and hypotheses to guide our investigation.

1. How well does ChatGPT-3-powered translation fare in Arabic-English machine translation?
2. Does ChatGPT-3 surpass Google Translate in its ability to translate between Arabic and English?
3. To what extent can ChatGPT be integrated into translation workflows particularly for Arabic-English translation tasks?

4. To what extent does ChatGPT-3's performance in Arabic-English machine translation align with or deviate from human translation proficiency?

The following hypotheses can be derived from the above research questions:

H1: ChatGPT-3-powered translation exhibits commendable performance in Arabic-English machine translation.

H2: ChatGPT-3 demonstrates superior performance in Arabic-English machine translation when compared to Google Translate.

H3: The integration of ChatGPT-3 into translation workflows has the potential to enhance the efficiency and effectiveness of Arabic-English translation tasks.

H4: The performance of ChatGPT-3 in Arabic-English machine translation partially aligns with human translation proficiency, yet do not fully replicate, the nuanced quality of translations achieved by human translators.

By addressing these research questions and testing the associated hypotheses, this study contributes to our understanding of ChatGPT's applicability for Arabic-English machine translation in an effort to contribute to this rapidly changing field.

II. Theoretical Background

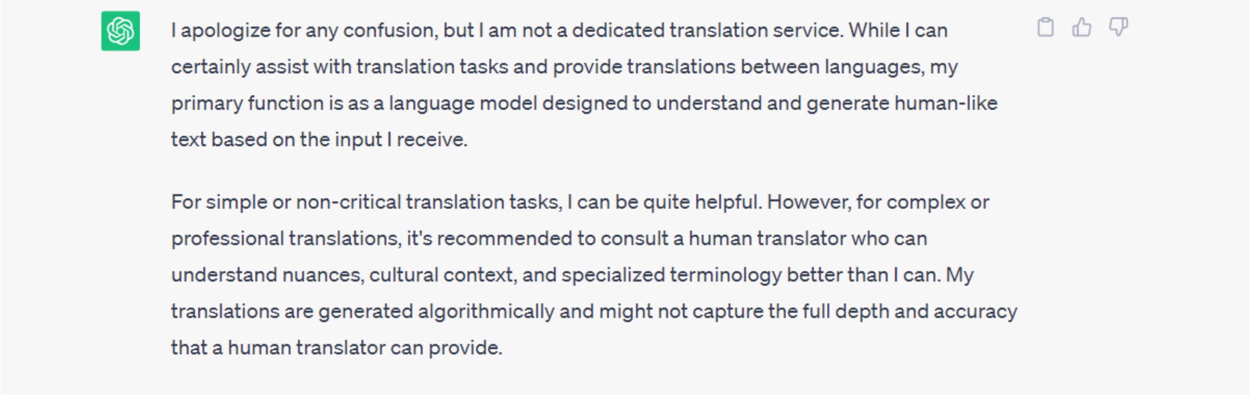
ChatGPT, a transformer-based model, utilizes natural language understanding to create translations. Unlike specialized machine translation systems, ChatGPT does not rely heavily on an extensive collection of parallel data for translation generation. Instead, it learns language structure and produces translations using a small collection of monolingual data (Frąckiewicz, 2023). The primary benefit of ChatGPT lies in its ability to generate high-quality translations even when provided with limited data. This characteristic makes it particularly suitable for languages with few available resources, a scenario often seen in low-resource languages (Frąckiewicz, 2023). Given that Arabic is regarded as a language with limited resources (Almansor et al., 2020; Amayreh & Amayreh, 2020), this research is significant since it examines ChatGPT's performance in translating between Arabic and English.

ChatGPT has proven its capacity to provide translations for low-resource languages, such as Vietnamese, Japanese and Korean, which are noticeably better in quality than those produced by specialized machine translation systems. This is mainly because ChatGPT adopts sophisticated deep learning technique to extract knowledge from sparse input, leading to more accurate translations (Nguyen et al., 2023). In short, ChatGPT is changing the way that low-resource languages are translated. Even with little data, ChatGPT can produce excellent translations due to its strong deep learning framework. This capability facilitates people's access to translations in their native languages (Frąckiewicz, 2023), representing a significant breakthrough in machine translation. Its influence on our approach to translation accessibility is bound to be enduring and transformative.

The initial launch of ChatGPT relied on the GPT-3.5 framework. The inclusion of "Chat" in its title alludes to its role as a conversational bot, while "GPT" is an abbreviation for generative pre-trained transformer, which denotes a category of extensive language models (LLM) (WEF, 2023). A new version, built upon GPT-4, the latest innovation from OpenAI, debuted on the 14th of March 2023. Access to this version is restricted to paying subscribers on a limited scale.

Though primarily designed as an intelligent conversational tool, ChatGPT can undertake various human-like tasks. such as crafting poems or resolving coding errors. Nevertheless, recent research by Jiao et al. (2023) indicates that ChatGPT, when provided with unsophisticated prompts, exhibits a noticeable disparity in performance when compared to other commercial translation systems like Google Translate and DeepL Translate. Unlike its counterparts, ChatGPT can adapt its output bias based on the prompt it receives. This means that users can input a range of translation prompts into the dialogue box alongside the source content rather than solely requesting translations from ChatGPT. Since OpenAI only offers a web interface for accessing ChatGPT, it is impossible to alter its internal components or access the intermediary representation of the system. Consequently, in the context of this research, we employed prompts for ChatGPT that delineated its translation task and contextual domain. This strategic approach served to guide its attention towards the specific input data, thereby aiming to augment the quality of the generated translations.

ChatGPT has been trained on extensive datasets encompassing a range of subjects. This prompts us to consider its proficiency in handling sentences tied to specific fields or potentially unclear ones. In order to activate ChatGPT's translation capabilities, we engaged in a conversation with the system. We inquired about its prowess in providing translation services and gauged its level of performance in this aspect. The answer is:



I apologize for any confusion, but I am not a dedicated translation service. While I can certainly assist with translation tasks and provide translations between languages, my primary function is as a language model designed to understand and generate human-like text based on the input I receive.

For simple or non-critical translation tasks, I can be quite helpful. However, for complex or professional translations, it's recommended to consult a human translator who can understand nuances, cultural context, and specialized terminology better than I can. My translations are generated algorithmically and might not capture the full depth and accuracy that a human translator can provide.

In addition, we provide prompts to ChatGPT in accordance with Gao et al. (2023), as prompts can improve ChatGPT's functionality. The researchers discovered that employing our suggested translation prompts can amplify ChatGPT's translation performance. The study's prompts encompass instructing ChatGPT to undertake Arabic-to-English translation tasks and identify the text's precise domain. For this research, we used the following prompt “Provide English translations for the following Arabic sentences. These sentences span various everyday subjects and are not limited to any particular genre or specialized field”.

According to the insights of Gao et al. (2023), when ChatGPT is furnished with accurate particulars concerning the input text, like the translation task and contextual domain through the prompt, it can significantly enhance its performance. However, if the information becomes exceedingly intricate or laden with noise, this could lead to substantial deterioration in performance. Moreover, incorporating the few-shot example technique warrants serious consideration as these instances carry a wealth of implicit information that cannot be explicitly conveyed through specific text passages.

According to information on the official website (OpenAI, 2023), ChatGPT has been refined using Reinforcement Learning from Human Feedback (RLHF). The aim is to boost the capabilities of these models enabling them to generate responses resembling human-like interactions. Through extensive exposure to vast amounts of text data, it learns to create detailed replies by adhering to instructions given in prompts. Despite its primary role as an intelligent conversational tool, ChatGPT is also adept at various human-like tasks, including machine translation. Nevertheless, recent research by Jiao et al. (2023) has highlighted a noticeable performance disparity between ChatGPT and other commercial translation systems like Google Translate and DeepL Translate. This disparity is even more pronounced when dealing with languages in short supply (Gao et al., 2023). As a result, we will delve into the comparative translation performance of ChatGPT and Google Translate using Arabic, a language with limited resources.

BLEU metric evaluation

Throughout the development of MT systems, a range of evaluation criteria are employed to gauge the enhancements made to these systems. These metrics can also serve the purpose of contrasting different MT systems. It is crucial to grasp the implications of the scores yielded by automated metrics when gauging the translation quality. These metrics are primarily built on the premise that MT's quality should resemble Human translation (HT). Human reference translation is a fundamental requirement to employ these automated metrics. Evaluating MT systems entails a comparative examination of their output concerning reference translation. These evaluation metrics furnish evaluative scores grounded in the reference translation that bears the closest resemblance (Rossi & Carr é, 2022).

The BLEU measurement assesses translation quality by considering adequacy and fluency. This involves evaluating how precisely words match between translations. The accuracy metrics like Precision, Recall, F-measure and BLEU-n are used to gauge the quality of translation. Higher scores in these metrics indicate better translation quality. For instance, BLEU focuses on the accuracy of n-grams. It measures how close the output from an MT system is to a professional HT of the same text. BLEU primarily distinguishes unsatisfactory and satisfactory MT outputs based on modified n-gram precision. This is calculated by comparing the matching n-grams between the translated text and source divided by the total n-grams in the translated text being evaluated. Each n-gram order's precision is calculated, and these precisions are then averaged together geometrically. In BLEU, the standard maximum n-gram order, frequently referred to as a string of four words, is employed. In order to discourage the use of sentences that are shorter than the reference translation, the metric calculates a modified precision score that has been adjusted with a brevity penalty. The resulting scores range from 0 to 1. Formulas 1 and 2 depict the calculation process employed by the BLEU metric. Formula 2 demonstrates how the BLEU score is calculated from the BP stated in Formula 1.

Formula 1

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

Formula 2

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

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

BLEU scores furnish a quantitative measure of translation quality by juxtaposing machine-generated translations against human-crafted counterparts. These scores (see figure 1) are assigned within a range that typically spans from 0 to 1, 0 to 10, or 0 to 100, with elevated values indicative of superior translation performance. In a simpler context, a perfect score of 100 signifies an impeccable alignment between the machine-generated translation and its human-authored counterpart (Aiken, 2019).

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

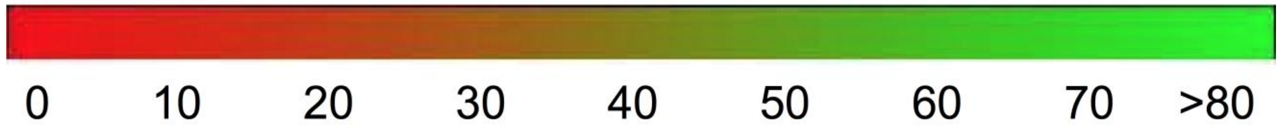


Figure 1. Interpretation of the BLEU Score (Evaluating Models | AutoML Translation Documentation | Google Cloud)

III. Literature Review

Recent advances in MT are primarily attributable to incorporating sizable language models like ChatGPT. This literature review aims to analyze and summarize findings from several studies exploring ChatGPT's capabilities, design approaches and evaluation methodologies in machine translation. Jiao et al. (2023) presented a compelling case for the enhanced translation capabilities of ChatGPT when powered by the GPT-4 engine. The study emphasized how GPT-4 integration significantly improved ChatGPT's translation capabilities, particularly in understanding idiomatic expressions, context, and finer linguistic details. The accuracy and fluency of the translations significantly improved as a result of this synergy. The study showed that ChatGPT, combined with the GPT-4 engine, became a proficient translator capable of competing with professional translation services for even far-off and low-resource languages (Jiao et al., 2023). The results confirmed ChatGPT's development into a skilled translator with expanded language coverage and improved robustness across various linguistic contexts, substantiating its expanding role in the machine translation landscape.

While Jiao et al. (2023) focused on the enhanced capabilities of ChatGPT with the GPT-4 engine, Gao et al. (2023) delved into the influence of prompt design on its translation quality. They (2023) focused on the influence of prompt design on ChatGPT's translation quality. Through empirical research, the researchers examined various translation prompts to improve the model's capacity to produce accurate and contextually appropriate translations. The study noted ChatGPT's potential to support machine translation and illuminated prompts' crucial role in maximizing translation outcomes. The research demonstrated that ChatGPT, when directed by carefully crafted prompts, could achieve comparable or even superior performance compared to commercial translation systems. This was done by experimenting with different prompts across various translations. The results showed how prompt design and ChatGPT's strength in natural language understanding and generation worked harmoniously, advancing machine translation capabilities.

Wang et al. (2023) explored the application of ChatGPT to document-level translation shedding light on its capacity to ensure coherence and consistency within extended texts. The study evaluated ChatGPT's performance in generating cohesive translations for entire documents providing crucial insights into its scalability within practical translation scenarios. Discourse modelling was a central theme in Wang et al.'s (2023) investigation, which focused on three main areas: the influence of Discourse-Aware Prompts on discourse phenomena and translation quality; a comparison of ChatGPT's translation abilities with advanced document-level MT methods and commercial MT systems and an analysis of discourse modelling capabilities and the effects of training techniques.

A comprehensive framework was presented by Peng et al. (2023) to utilize ChatGPT's capabilities in machine translation. They proposed strategies encompassing input text preprocessing, model fine-tuning and output post-processing to enhance translation accuracy and coherence. The study discussed the drawbacks of earlier methods that relied on straightforward prompts and fell short of utilizing ChatGPT's capabilities entirely. Domain-Specific Prompts

(DSP) and Task-Specific Prompts (TSP) were developed due to the researchers' investigations into several variables, including temperature, task information and domain information. The results provided key insights: lowering temperature improved performance, task emphasis improved results, domain knowledge improved generalization, and these prompts could reduce ChatGPT's propensity for hallucinations in non-English-centric MT tasks.

In contrast to Peng et al. (2023), who presented strategies to enhance translation accuracy, the study by Bang et al. (2023) explored a wider context of ChatGPT's reasoning abilities and interactivity. Bang et al.'s (2023) conducted thorough analysis that examined ChatGPT in a wider context. It included a range of tasks, languages and modalities. This study went beyond translation evaluation to explore ChatGPT's reasoning abilities, resistance to hallucinations, and interactivity in translation scenarios. The authors suggested a quantitative evaluation framework for interactive Large Language Models (LLMs) like ChatGPT using 23 datasets covering eight different NLP application tasks. This evaluation considered multitasking, multilingualism, and multimodal interaction by combining existing datasets with a newly created multimodal dataset. The study's results demonstrated ChatGPT's impressive performance, frequently outperforming zero-shot learning benchmarks and even refined models on various tasks.

While Peng et al. (2023) and Bang et al. (2023) presented comprehensive frameworks and assessed ChatGPT's performance across a range of tasks and modalities, the research understudy focused on the translation aspect and its application to a particular language pair. It focused on Arabic-English machine translation and contrasted ChatGPT's performance against Google Translate. By doing so, we address a research gap by examining ChatGPT's potential in this field, while also advancing our understanding of its overall performance.

The study by Lu et al. (2023) introduced a novel method called "Error Analysis Prompting" to assess the translation quality of LLMs, such as ChatGPT, in a way that closely resembles human evaluation. This method involved identifying specific error types in the translation output and prompting the model to rectify these errors. Error Analysis Prompting enables LLMs such as ChatGPT to produce MT evaluations that resemble human evaluations at the system and segment levels. The research findings not only discussed ChatGPT's advantages and disadvantages as an MT evaluator but also offered insightful advice on how to create powerful prompts for context-aware learning, encouraging further advancements in translation assessment methods.

Similarly, Lyu, Xu and Wang (2023) explored MT's dynamic landscape facilitated by LLMs like ChatGPT. They explained how ChatGPT was applied in novel ways in MT to address the changing problems in the field. Their study effectively pushed the boundaries of translation quality and efficiency in the context of MT by utilizing ChatGPT's capabilities. The authors envisioned and suggested various futures for MT based on LLMs, such as stylized MT, interactive MT and MT based on Translation Memory. These novel directions highlighted the potential of LLMs and herald the emergence of cutting-edge strategies to improve translational effectiveness.

The performance of ChatGPT, specifically in translating between Arabic and English, was examined in the study by Khoshafah (2023). The study aimed to evaluate the ChatGPT 3.5 model's translation precision in light of its cross-cultural communication applications. The study entailed translating texts of various genres, including those with historical, literary, media, legal and scientific content and then contrasting ChatGPT's translations with those of trained human translators. Language accuracy and context coherence were the evaluation criteria. The results highlighted ChatGPT's aptness for translating simpler content and its limitations in accurately comprehending complex texts.

Despite the thorough exploration of numerous aspects of ChatGPT's capabilities and its application in machine translation within previous literature analysis, this study distinguishes itself by directly comparing the performance of ChatGPT-3 and Google Translate with a specific focus on translating from Arabic to English. This comprehensive comparison sheds light on the performance of these two well-known translation systems.

Furthermore, the assessment of their performance not only grants us valuable insights into their suitability for seamless integration into diverse workflows but also underscores the far-reaching implications and practical applications of incorporating ChatGPT into contexts entailing professional translation. Moreover, the appraisal of ChatGPT-3's performance in relation to the proficiency of human translators, as evaluated through the BLEU metric, serves to enrich our understanding of machine translation technology's capacity to achieve translation quality on par with that produced by humans.

IV. Methodology

This research adopts a quantitative research methodology. The process includes systematic data gathering, thorough analysis, and rigorous evaluation of translation quality within the established context of recognized machine translation metrics and standards.

A. Data Collection Procedures

Due to the constraints posed by the limited availability of Arabic datasets and computing resources, our data collection process focused on a subset of 1000 arbitrary sentences. These sentences were extracted from Tatoeba (2023), an online collaborative platform that offers an extensive array of sentences in numerous languages. Tatoeba's main objective is to establish a varied and all-encompassing repository of sentences beneficial for acquiring languages, practicing translation and conducting linguistic investigations. Tatoeba operates as an open-source initiative, with its data accessible through a Creative Commons license. Consequently, the material on Tatoeba can be freely employed and distributed contingent upon proper acknowledgement and adherence to the licensing conditions.

To ensure methodological rigor, both ChatGPT and Google Translate were accessed via their publicly available web interfaces. A common set of sentences was selected for translation using both systems, ensuring the uniformity of evaluation conditions and minimizing the influence of random variables. This meticulous approach enhances the reliability and comparability of our analysis.

The choice to employ the free versions of Google Translate and ChatGPT aligns seamlessly with prevailing market trends. A significant proportion of users, spanning individuals to small enterprises, opt for these accessible solutions to meet their translation needs (Li et al., 2020; Alves, 2019; Plitt & O'Brien, 2018; Koehn et al., 2017). The affordability and ease of access offered by these free versions make them particularly suitable for a comprehensive investigation of this nature.

In line with best practices, Google Translate was chosen as a benchmark for comparison due to its generally reliable performance. Established in 2006, Google Translate has grown to become one of the leading machine translation tools. It currently supports 133 languages, with an additional 24 languages added in 2022 (Harby, 2023). The level of accuracy varies depending on the language combination and the content being translated. Some reports suggest that Google Translate can achieve an accuracy rate of up to 94% (Castilho et al., 2019). The pivotal moment in Google's translation quality came in 2016 when it adopted NMT. This transition significantly improved the quality of translated output. According to the information provided by the technology giant, Google Neural Machine Translation (GNMT) reduced over 60% of translation errors for major language pairs (Harby, 2023). Regarding dependability and precision, Google Translate stands out, particularly for languages with limited resources. In 2022, Google Translate secured the top position among 18 other machine translation engines for nearly all language combinations, as revealed by the MT evaluation program conducted by Intento (Harby, 2023).

B. Data Analysis

To gauge the performance of both ChatGPT and Google Translate, we employ the BLEU metric, a cornerstone of machine translation evaluation. While alternative metrics such as METEOR, TER, BLEU, and NIST exist within the machine translation community, the prevailing consensus endorses BLEU as the primary yardstick for developers (Maučec & Donaj, 2019). As attested by Warner (2022), BLEU stands as the most frequently adopted metric, underscoring its relevance and widespread usage. Accordingly, BLEU scores constitute the focal point of assessment within our study, affording a comprehensive view of translation precision.

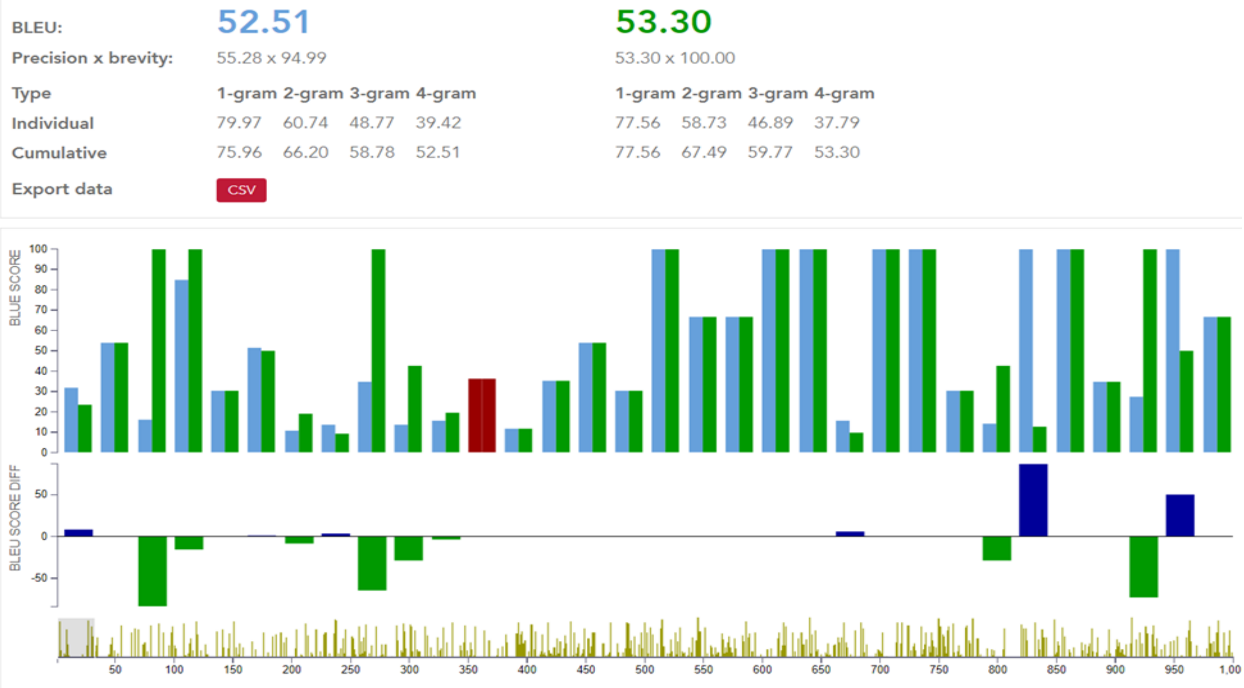
Our decision to employ BLEU scores aligns with the aim of capturing the nuanced differences between the translations generated by ChatGPT and Google Translate. This metric not only provides an objective and consistent means of assessment but also grants us the ability to delineate the varying degrees of fidelity each system achieves when translating from Arabic to English. As we delve into the ensuing analysis, these BLEU scores will serve as the bedrock upon which the comparative evaluation of translation accuracy will be conducted.

V. Results and Discussion

The analysis focused on comparing the ratio of flawed sentences to those lacking meaningful insights in machine translation. The findings were presented in figure 2 which displays the outcomes of an automated calculation determining the accuracy of ChatGPT and Google MT systems across varying four-gram sizes. The process entailed computing the BP, identifying the reference with more common n-grams, computing the length (represented by 'r' as per formula 1) and then evaluating the total length of the MT translation denoted as 'c'. This method amalgamated precision values into an overall score called the BLEU score.

After evaluating 1000 sentences in our dataset, the n gram scores for ChatGPT (green) were found to be as follows;1-gram: 77.56, 2-gram: 67.49, 3-gram: 59.77, 4-gram: 53.30. These scores were found to surpass those of Google Translate (blue) which were: 1-gram: 75.96, 2-gram: 66.20, 3-gram: 58.78, 4-gram: 52.51".

The results signify a more favorable outcome for ChatGPT, indicating that it produced translations with a higher degree of accuracy and meaningful content when compared to Google Translate. Specifically, ChatGPT achieved a BLEU score of 53.30, surpassing Google Translate's score of 52.51. However, it is important to note that even with ChatGPT's more positive performance, the analysis still revealed the presence of errors in a number of examined sentences, suggesting the need for subsequent editing to enhance translation quality and overall coherence.



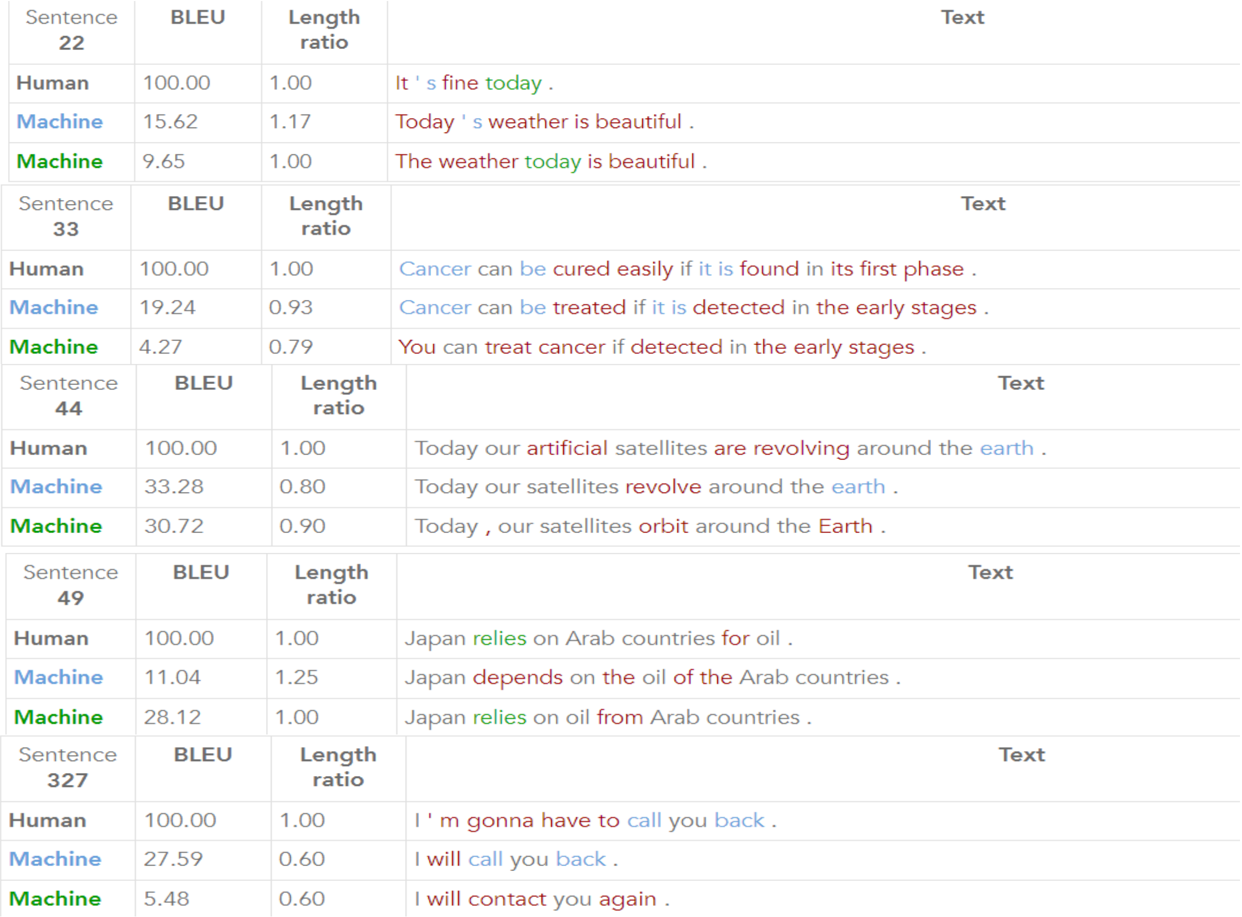
BLEU:	52.51				53.30			
Precision x brevity:	55.28 x 94.99				53.30 x 100.00			
Type	1-gram	2-gram	3-gram	4-gram	1-gram	2-gram	3-gram	4-gram
Individual	79.97	60.74	48.77	39.42	77.56	58.73	46.89	37.79
Cumulative	75.96	66.20	58.78	52.51	77.56	67.49	59.77	53.30
Export data	CSV							

Figure 2. BLEU Scores for Google Translate and ChatGPT

🟦 Google Translate

🟩 ChatGPT

To illustrate the observed variations in translation quality between ChatGPT (green) and Google Translate (blue), the following examples are taken randomly through the BLEU metric system.



Sentence 22	BLEU	Length ratio	Text
Human	100.00	1.00	It ' s fine today .
Machine	15.62	1.17	Today ' s weather is beautiful .
Machine	9.65	1.00	The weather today is beautiful .

Sentence 33	BLEU	Length ratio	Text
Human	100.00	1.00	Cancer can be cured easily if it is found in its first phase .
Machine	19.24	0.93	Cancer can be treated if it is detected in the early stages .
Machine	4.27	0.79	You can treat cancer if detected in the early stages .

Sentence 44	BLEU	Length ratio	Text
Human	100.00	1.00	Today our artificial satellites are revolving around the earth .
Machine	33.28	0.80	Today our satellites revolve around the earth .
Machine	30.72	0.90	Today , our satellites orbit around the Earth .

Sentence 49	BLEU	Length ratio	Text
Human	100.00	1.00	Japan relies on Arab countries for oil .
Machine	11.04	1.25	Japan depends on the oil of the Arab countries .
Machine	28.12	1.00	Japan relies on oil from Arab countries .

Sentence 327	BLEU	Length ratio	Text
Human	100.00	1.00	I ' m gonna have to call you back .
Machine	27.59	0.60	I will call you back .
Machine	5.48	0.60	I will contact you again .

Figure 3 (a, b, c, d, e). Examples of Translation Quality Comparison
Randomly Selected Examples of Translations From the Dataset, Evaluated Using the BLEU Metric System
Source: Adapted From Data Collected in This Study.

This paper aimed to evaluate ChatGPT's performance in machine translation tasks, specifically comparing its performance with that of Google Translate. For this purpose, the following research questions were formulated:

1. How well does ChatGPT-3-powered translation fare in Arabic-English machine translation?
2. Does ChatGPT-3 surpass Google Translate in its ability to translate between Arabic and English?
3. To what extent can ChatGPT be integrated into translation workflows particularly for Arabic-English translation tasks?
4. To what extent does ChatGPT-3's performance in Arabic-English machine translation align with or deviate from human translation proficiency?

Based on the analysis, we can now discuss the findings of this research in relation to the abovementioned research questions and the corresponding hypotheses.

H1: ChatGPT-3-powered translation exhibits commendable performance in Arabic-English machine translation.

We confirmed the research hypothesis as our analysis demonstrates that ChatGPT-3-powered translation displays commendable performance in the realm of Arabic-English machine translation. Based on the interpretation of BLEU scores in figure 1, ChatGPT's score signifies the translation's high quality. The model categorized the outcome as "Very high quality, sufficient, and fluent translations".

H2: ChatGPT-3 demonstrates superior performance in Arabic-English machine translation when compared to Google Translate.

We confirmed the research hypothesis as BLUE score for ChatGPT is higher than Google Translate score.

H3: The integration of ChatGPT-3 into translation workflows has the potential to enhance the efficiency and effectiveness of Arabic-English translation tasks.

We confirmed the research hypothesis as according to figure 1, the model categorized the score as "Very high quality, sufficient, and fluent translations." This implies that it can be seamlessly integrated into professional workflows.

H4: The performance of ChatGPT-3 in Arabic-English machine translation partially aligns with human translation proficiency, yet does not fully replicate the nuanced quality of translations achieved by human translators.

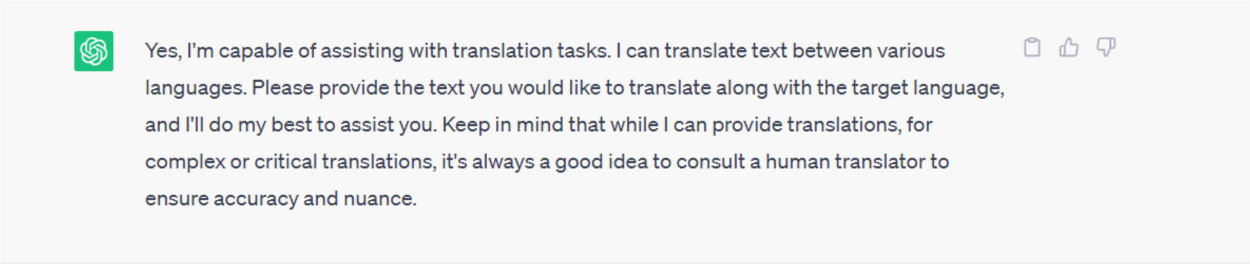
We confirmed the research hypothesis as the results demonstrated that ChatGPT-3's performance in Arabic-English machine translation is “Very high quality, sufficient, and fluent translations” according to the BLEU score of 53.30. However, this score remains below the threshold of surpassing human translation quality (typically above 60), indicating a partial replication of the nuanced and contextually rich translations achieved by human translators.

By addressing these research questions and testing the associated hypotheses, this study offers fresh perspectives on the practical utility, comparative effectiveness, and human parity elements of ChatGPT-3 in the context of Arabic-English translation, and, thus, contributes to the advancement of academic discourse. This multifaceted exploration contributes to the theoretical discourse and carries significant implications for the practical implementation of these translation systems in real-world scenarios.

Based on our analysis, one could argue that ChatGPT can be used as long as it is not your only step in the translation workflow. It can be used in the translation workflow if additional steps are taken; grammatical and spelling proofreading and localization for cultural nuances.

ChatGPT could perform the same tasks as other machine translation tools. It assists translators in providing even better, faster, and more efficient translations of spoken or written text. It reduces the duration and expense of translation services enabling translators to produce translations more quickly without compromising quality. Also, it reduces the cost and increases the availability of translation services, particularly for individuals and small businesses.

When we asked ChatGPT if it was good at translating languages, the tool admitted its shortcomings. In the machine's words:



Yes, I'm capable of assisting with translation tasks. I can translate text between various languages. Please provide the text you would like to translate along with the target language, and I'll do my best to assist you. Keep in mind that while I can provide translations, for complex or critical translations, it's always a good idea to consult a human translator to ensure accuracy and nuance.

The ongoing debate surrounding machine versus HT persists, revolving around whether MT will eventually supplant HT, particularly at a time when MT continues to make strides. MT has notably diminished the language barrier. After all, MT surpasses humans in at least two key aspects of translation: its ability to work much more swiftly and cost-effectively. These advantages are particularly appealing in today's landscape, where saving time and money is paramount for most businesses. Consequently, certain translators express apprehension that excessive progress in MT might jeopardize their professional prospects.

However, MTs are riddled with numerous shortcomings, limiting their applicability in various spheres of life. Outputs from platforms like Google Translate, ChatGPT-3 powered translation and comparable systems only serve a specific purpose: to glean the overall meaning of the source text. However, human ingenuity and intellect remain integral components of translation, and thus far, no software has succeeded in replicating these qualities.

MT is and will continue to be utilized, but the necessity for human evaluation of translation quality persists if only to ensure accuracy (Puchała-Ladzińska, 2016). Machines can expedite the translation process, yet they can neither entirely replace the human factor nor attain the pinnacle of excellence.

In the meantime, machine translation systems should be regarded as tools for translation assistance, while human translators take on the role of post-editors. Machine translation can establish a base for professional translators to review, rephrase, enhance the writing style and adapt the content to fit the specific context and audience of the target language. This means that instead of starting from scratch, the translator cross-checks, proofreads and refines machine-generated translations. A notable advantage of this collaboration between humans and machines is the increased efficiency of the translator.

The connection between machines and humans is one of mutual supplementation. According to statistics and ongoing research, modern technologies like machine translation will never be capable of completely supplanting humans; their role is more about aiding rather than jeopardizing human translators (Şahin & Gürses, 2021). Translators adept in using machine translation possess a competitive edge over those unfamiliar with current translation tools.

Hence, the apprehensions among human translators regarding potential displacement by machines in the future are unfounded. Nevertheless, the role of the translator is anticipated to evolve inevitably. Human translators might shift from being direct translators to becoming editors, refining materials previously translated by machines as machine translation systems advance further.

VI. Conclusion

This study has preliminarily assessed ChatGPT's capabilities for machine translation between Arabic and English. This study aimed to evaluate the translation capabilities of ChatGPT and identified its position in the translation workflow by comparing its output with that of Google Translate. The study used a comparable corpus of 1000 Tatoeba sentences assessed using the BLEU metric. The results showed that ChatGPT outperforms commercial translation systems like Google Translate in producing high-quality output and even slightly outperforms them when given specific translation tasks and context domains. Despite significant progress in ChatGPT's translation capabilities, the study emphasized the value of human translators in preserving translation quality, particularly for complex and nuanced content. The coexistence of human expertise and machine translation tools was recommended to enhance overall translation accuracy and fluency. Further research opportunities lie in exploring ChatGPT's performance in other language pairs assessing its fluency and naturalness, and conducting human evaluations to provide deeper insights into its strengths and limitations. Machine translation tools like ChatGPT are valuable resources in the dynamic machine translation market; however, they cannot replace the originality and nuanced thinking that human translators bring to the table. The future likely involves a collaborative approach that leverages the strengths of both humans and machines to achieve optimal translation outcomes.

The importance of this study is found in its in-depth analysis of ChatGPT's performance in the context of Arabic-English translation. We emphasize ChatGPT's potential benefits and shed light on its function as a tool for improving Arabic-English translation tasks by conducting a methodical evaluation and benchmarking against Google Translate. Our study of ChatGPT's incorporation into translation workflows further advances the technology's practical application by highlighting its usefulness in actual situations. By conducting this study, we hope to advance knowledge of machine translation's capabilities and limits, particularly in the context of Arabic-English translation, and to help decision-makers make well-informed choices when integrating AI technologies into translation workflows.

The outcomes of this study have the potential to illuminate the variations in translation quality between human and machine translation. This investigation carries significance as it lays the foundation for a theoretical framework concerning the accuracy of machine translation. In a broader context, there is a genuine effort to enhance our comprehension of the effectiveness of machine translation compared to HT in converting Arabic to English texts. This holds value for translators, students, educators and specialists in the field. The insights gained from this research can be advantageous for software developers working in the machine translation domain, aiding them in enhancing the quality of machine-generated translations. Additionally, these findings may serve as a valuable resource for field experts engaging in comparative analyses within machine versus HT.

While this study concentrated on translating Arabic source text into English target text, extending research to encompass other language pairs is worth considering. Furthermore, this current investigation was confined to short texts underscoring the recommendation for future research to include longer passages to establish wider applicability. Similarly, the scope of this work was restricted to just two machine translation systems: ChatGPT3 powered translation and Google Translate. Subsequent research could delve into other machine translation systems, linguistic aspects, a wider variety of texts and human translators.

REFERENCES

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

[2] Almansor, E.H., Al-Ani, A., Hussain, F.K. (2020). Transferring Informal Text in Arabic as Low Resource Languages: State-of-the-Art and Future Research Directions. In: Barolli, L., Hussain, F., Ikeda, M. (eds) Complex, Intelligent, and Software Intensive Systems. CISIS 2019. *Advances in Intelligent Systems and Computing*, *vol 993*. Springer, Cham. https://doi.org/10.1007/978-3-030-22354-0_17

[3] Alves, F., Ferreira, A., & Pinto, J. (2019). The Use of Machine Translation in Small and Medium-Sized Enterprises: A case Study. *Journal of Business Research*, *104*, p. 292-300.

[4] Amayreh, H. & Amayreh, M. (2020). Linguistic Branching of Semantics in Arabic: a Social approach. *Al-Balqa Journal for Research and Studies*, *23*(2), 147-156. 10.35875/1105-023-002-012

[5] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W. et al. (2023). *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*. DOI: 10.48550/arXiv.2302.04023

[6] Brown, T. Mann, b., Ryder, N. et al. (2020). Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, *33*, pp. 1877–1901.

[7] Chen, M., Tworek, J., Jun, H. et al. (2021). *Evaluating Large Language Models Trained on Code*. https://doi.org/10.48550/arXiv.2107.03374.

[8] Dai, H., Xhafa, F., Janse, B.J., LIang, H. and Ye, J. (2022). Comparative Analysis of Machine Translation and Human Translation under the Background of the Internet. *International Conference on Cognitive-based Information Processing and Applications (CIPA 2021)*, *1*(84), pp. 877-882.

[9] Frąckiewicz, M. (2023). ChatGPT's Contributions to Improving Translation Quality for Low-Resource Languages. Retrieved July 26, 2023 from Artificial intelligence. *ChatGPT's Contributions to Improving Translation Quality for Low-Resource Languages* (ts2.space).

[10] Gao, Y., Wang, R. and Hou, F. (2023). *How to Design Translation Prompts for ChatGPT: An Empirical Study*. *arXiv e-prints*, pp.arXiv-2304. https://doi.org/10.48550/arXiv.2304.02182

[11] Gu, W. (2023). *Linguistically Informed ChatGPT Prompts to Enhance Japanese-Chinese Machine Translation: A Case Study on Attributive Clauses*. https://doi.org/10.48550/arXiv.2303.15587.

[12] Harby, A. (2023). How Accurate is Google Translate? Retrieved from: *How Accurate is Google Translate?* - Slator

[13] He, J., Neubig, G. and Berg- Kirkpatrick, T. (2021). Efficient Nearest Neighbor Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5703–5714.

[14] Jiao, W., Wang, W., Huang, J.T., Wang, X. and Tu, Z.P. (2023). *Is ChatGPT a Good Translator? Yes with GPT-4 as the Engine*. Retrieved June 14, 2023 from https://www.researchgate.net/publication/367359399_Is_ChatGPT_A_Good_Translator_A_Preliminary_Study.

[15] Khoshafah, F. (2023). *ChatGPT for Arabic-English Translation: Evaluating the Accuracy*. DOI: 10.21203/rs.3.rs-2814154/v1

[16] Kocmi, T. and Federmann, C. (2023). *Large Language Models Are State-of-the-Art Evaluators of Translation Quality*. Retrieved June 16, 2023 from 2302.14520.pdf (arxiv.org).

[17] Koehn, P., Hoang, H., & Birch, A. (2017). *The Future of Machine Translation: A Survey of Industry Experts*. arXiv preprint arXiv:1707.03814.

[18] Li, J., Wang, J., & Wang, H. (2020). The Adoption of Machine Translation by Individuals and Small-Scale Businesses: A survey. *Computers in Human Behavior*, *111*, 106293.

[19] Li, X. and Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL 2021*.

[20] Li, Y., Yin, Y., Li, J. and Zhang, Y. (2022). Prompt-Driven Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL*.

[21] Liu, Y. Ott, M. Goyal, N. Du, J. Joshi, M. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*, arXiv:1907.11692. Retrieved October 10, 2023 from doi: https://doi.org/10.48550/arXiv.1907.11692

[22] Lu, Q., Qiu, B., Ding, L., Xie, L. and Tao, D. (2023). *Error Analysis Prompting Enables Human-Like Translation Evaluation In Large Language Models: A Case Study On ChatGPT*. arXiv preprint arXiv:2303.13809.

[23] Lyu, C., Xu, J. and Wang, L. (2023). *New Trends in Machine Translation Using Large Language Models: Case Examples with ChatGPT*. arXiv preprint arXiv:2305.01181.

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

[25] Nguyen, T. T. H., Nguyen, V. H., Nguyen, T. A., & Nguyen, P. T. (2023). *Using ChatGPT for Machine Translation Between Languages without Parallel Data*. arXiv preprint arXiv:2302.00002.

[26] *OpenAI*. (2023). Retrieved August 14, 2023 from: ChatGPT (openai.com).

[27] Ouyang, L., Wu, J., Jiang, X. et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*. Retrieved July 16, 2023 from b1efde53be364a73914f58805a001731-Paper-Conference.pdf (neurips.cc)

[28] Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y. and Tao, D. (2023). *Towards Making the Most of ChatGPT for Machine Translation*. arXiv preprint arXiv:2303.13780.

[29] Plitt, M., & O'Brien, S. (2018). The Impact of Machine Translation on the Translation Industry. *Journal of Translation Studies*, *21*(2), 117-134.

[30] Puchała-Ladzińska, K. (2016). Machine Translation: A Threat or an Opportunity for Human Translators?", *Studia Anglica Resoviensia T*, *3*, pp. 89-98, doi: 10.15584/sar.2016.13.9.

[31] Rossi, C. & Carr é, A. (2022). How to Choose a Suitable Neural Machine Translation Solution: Evaluation of MT quality. In Dorothy Kenny (ed.), *Machine Translation for Everyone: Empowering Users In The Age of Artificial Intelligence*, 51–79. Berlin: Language Science.

[32] Şahin, M. and G ürses, S. (2021). English-Turkish Literary Translation Through Human-Machine Interaction. *Revista Tradum àtica*, *19*, pp. 306-310. doi: 10.5565/rev/tradumatica.284.

[33] Stahlberg, F. (2020). Neural Machine Translation: A review. *Journal of Artificial Intelligence Research*, *69*(2020), pp. 343-418.

[34] *Tatoeba*. (2023). Retrieved August 2, 2023 from: https://tatoeba.org/

[35] Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). *Document-Level Machine Translation With Large Language Models*. arXiv preprint arXiv:2304.02210.

[36] Warner, A. (2022). *Humans Still Beat Machines When It Comes to Literary Translation*. Retrieved from: MultiLingual.

[37] Wei, J., Bosma, M., Zhao, V., Guu, K. et al. (2022). Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*. Retrieved July 14, 2023 from pdf (openreview.net).

[38] World Economic Forum (WEF). (2023). Retrieved July 16, 2023 from *Generative AI – A Game-Changer Society and Industry Needs to be Ready for | World Economic Forum* (weforum.org).

Linda Alkhawaja is an Associate Professor in the Department of Translation Studies at Al-Ahliyya Amman University in Jordan. She holds an MA in Translation and Interpreting Studies from Salford University/Manchester and a PhD in Sociology in Translation Studies from Aston University/Birmingham. She is actively engaged in various projects exploring the sociology of translation, cultural studies, and Machine Translation. Her work delves into the complex relationship between language, culture, and society, unraveling the transformative power of translation in different contexts. She took on, for two years, the role of the Director of the Language Centre and the Head of the English Language and Literature and English Language and Translation Departments. She recently developed a keen interest in artificial intelligence (AI). She undertook an accredited online diploma program in the fundamentals of artificial intelligence, certified by the Continuing Professional Development (CPD), adding a new dimension to her academic journey. Her research now bridges the worlds of technology and AI with the field of translation studies.