Human-Centric Intelligent Systems
https://doi.org/10.1007/s44230-024-00061-7

RESEARCH ARTICLE

Error Analysis of Pretrained Language Models (PLMs) in English-to-Arabic Machine Translation

Hend Al-Khalifa[1,3] · **Khaloud Al-Khalefah**[2] · **Hesham Haroon**[3]

Received: 3 October 2023 / Accepted: 4 January 2024

Abstract
Advances in neural machine translation utilizing pretrained language models (PLMs) have shown promise in improving the translation quality between diverse languages. However, translation from English to languages with complex morphology, such as Arabic, remains challenging. This study investigated the prevailing error patterns of state-of-the-art PLMs when translating from English to Arabic across different text domains. Through empirical analysis using automatic metrics (chrF, BERTScore, COMET) and manual evaluation with the Multidimensional Quality Metrics (MQM) framework, we compared Google Translate and five PLMs (Helsinki, Marefa, Facebook, GPT-3.5-turbo, and GPT-4). Key findings provide valuable insights into current PLM limitations in handling aspects of Arabic grammar and vocabulary while also informing future improvements for advancing English–Arabic machine translation capabilities and accessibility.

Keywords Machine translation · Pretrained large language models · Translation studies · GPT · Arabic language

1 Introduction

The digital era has brought about a significant shift in communication and interaction across linguistic and cultural boundaries. With the rise of global interconnectedness, the ability to translate between languages, particularly those using non-Latin scripts, such as Arabic, has become increasingly important. However, many writing systems beyond the Latin alphabet face barriers to digital accessibility and participation [1]. Artificial Intelligence (AI), particularly in the field of Neural Machine Translation (NMT), has shown promise in bridging these linguistic divides [2, 3]. However, translating between languages with profound structural and script differences, such as English and Arabic, presents unique challenges [4, 5].

Pretrained Language Models (PLMs) are at the forefront of Natural Language Processing (NLP) research and have significantly enhanced machine translation capabilities [6]. Despite their advancements, these models still struggle to achieve high levels of accuracy and fluency in English-to-Arabic translation [7]. This limitation hampers effective communication and cooperation across English and Arabic-speaking cultures, which is crucial in areas such as trade, diplomacy, and knowledge exchange. This study aims to investigate the error patterns in state-of-the-art PLMs when translating from English to Arabic. Our objectives will be:

- To evaluate these models in a parallel corpus of English–Arabic sentences.
- To identify common error patterns and explore their possible causes.
- To provide insights into the limitations of current PLMs.

This study targets academics, NLP practitioners, and policymakers interested in leveraging AI for language translation. Our findings provide valuable insights into the current limitations of PLMs in terms of English-to-Arabic translation quality. The analysis intends to offer guidance for future research efforts focused on advancing machine translation capabilities for the English–Arabic language pairs. Additionally, by elucidating the existing challenges in cross-lingual

✉ Hend Al-Khalifa
hendk@ksu.edu.sa

Khaloud Al-Khalefah
kholodalkhalifah@gmail.com

[1] Information Technology Department, King Saud University, Riyadh, Saudi Arabia

[2] College of Language and Translation, Al-Imam Muhammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

[3] iWAN Research Group, King Saud University, Riyadh, Saudi Arabia

Published online: 05 February 2024

AI systems, we hope to highlight the broader significance of progress in this space to achieve equitable global participation in the exchange of ideas.

The rest of the paper is organized as follows: Sect. 2 provides background and related work, discussing Pretrained Language Models, English-to-Arabic machine translation, and error analysis. Section 3 details the methodology, including the dataset, PLM selection and training, and the evaluation metrics used. Section 4 presents the results and analysis, comprising a model performance comparison, error classification, and patterns. Section 5 concludes the paper by summarizing the key findings and implications for improving English-to-Arabic translation. Finally, Sect. 6 offers recommendations and future work, highlighting potential avenues for enhancing the PLM performance in this language pair.

2 Background and Related Work

Machine translation, pioneered in the 1950s, initially saw success with the Georgetown-IBM experiment, utilizing statistical algorithms for Russian-to-English translations. Over time, this field has diversified into rule-based systems, statistical approaches, neural networks, and Large Language Models, each with unique methods for learning and translating languages. These systems, especially neural networks and LLMs, continuously improve their translation accuracy by training on extensive datasets and grasping the underlying knowledge in the text [8].

In this section, we briefly discuss three topics related to our research: English-to-Arabic machine translation research, Pretrained Language Models (PLMs) in machine translation, and Error Analysis in machine translation.

2.1 English-to-Arabic Machine Translation

Machine translation from English to Arabic has been an active area of research for several years. Researchers have explored various approaches to tackling the challenges inherent in this task, including differences in grammar, word order, and vocabulary between the two languages. In addition, many survey papers have been published on Arabic linguistic characteristics and translation challenges. For instance, Ameur et al. [9] summarized critical research on Arabic MT and the available tools/resources for building Arabic MT systems. The survey discussed the state of the field and provided insights into future Arabic MT research directions. In addition, Zakraoui et al. [10] provided a comprehensive review comparing different NMT approaches for Arabic-English translations. They discussed approaches addressing linguistic and technical challenges, and demonstrated success over traditional methods. Their results will serve to update researchers on resources for improving Arabic MT, including corpora, toolkits, techniques, and models.

Rule-based methods, which rely on handcrafted rules to analyze the source language and generate the target language, have shown some success. Farhat and Al-Taani [11] developed a rule-based system that could translate simple English sentences into Arabic with an accuracy of 85.71%. Similarly, Alawneh et al. [12] combined rule-based and example-based English-to-Arabic machine translation using parsing and a hybrid methodology to handle ordering and agreement. They evaluated their approach on 250 test samples, and the results achieved 97.2% precision on average. Also, Al-Rukban and Saudagar [13] evaluated three commercial English-to-Arabic systems, Google Translate, Bing Translator, and Golden Alwafi, and found that Golden Alwafi achieved the highest BLEU score, indicating the most human-like translations. Although rule-based methods are straightforward, they require extensive time to develop and maintain.

Neural-driven methods, including those based on neural networks and PLMs, have become increasingly popular. Akeel and Mishra [14] developed an English-to-Arabic translator using both rule-based and neural network methods, achieving scores of 0.6029 on the n-gram BLUE score and 0.8221 on the METEOR metric. Aljohany et al. [15] proposed a bidirectional model for the translation between Arabic and English. This model employs a Long Short-Term Memory (LSTM) encoder-decoder with an attention mechanism to address the performance degradation linked to increased input sentence length. The integration of LSTM and attention mechanisms improves the translation accuracy, as substantiated by the experimental results, which demonstrate improved translation precision and reduced loss. Some researchers have focused on the challenges of English-to-Arabic translation and have suggested directions for future work. Aref et al. [16] outlined a multi-level approach to machine translation and reviewed the state of English-to-Arabic translation, suggesting the use of AI techniques like knowledge representation to build a prototype system. In contrast, Nagoudi et al. [17] developed TURJUMAN, a toolkit leveraging the Transformer AraT5 model, and translated 20 languages into Modern Standard Arabic (MSA).

Table 1 summarizes the main rule-based and neural approaches explored for English-to-Arabic machine translation. For each approach, key published works are highlighted, along with their main findings, limitations, and open gaps in the research. This provides a concise overview of the current state of English-to-Arabic MT literature to identify promising future research directions. We can see that the key gaps in English-to-Arabic translation include insufficient parallel data, challenges with Arabic morphology, and linguistic divergence. Our research utilizes a new

Table 1 Summary of key approaches, findings, limitations, and open gaps in English-to-Arabic machine translation literature

Approach	Key works	Findings	Limitations
Rule- based	Farhat and Al-Taani [11]: Rule-based system for simple English-to-Arabic translation, 85.71% accuracy Alawneh et al. [12]: Combined rule-based and example-based approach, 97.2% precision Al-Rukban and Saudagar [13]: Evaluated commercial systems, Golden Alwafi had highest BLEU	Can achieve good accuracy and precision for simple sentences Requires extensive manual effort for rules	Do not scale well to complex sentences Hard to maintain rules over time
Neural methods	Akeel and Mishra [14]: Combined rule-based and neural, BLEU of 0.6029 Aljohany et al. [15]: LSTM encoder-decoder with attention for better long sentence translation Nagoudi et al. [17]: TURJUMAN toolkit with Transformer model	Neural models outperform rule-based Attention mechanisms help with long sentences	Limited focus on English–Arabic specifically More complex linguistic challenges not fully solved

English–Arabic parallel corpus and several PLMs models on this data to adapt it and assess the translation performance.

2.2 Pretrained Language Models (PLMs) in Machine Translation

Pretrained language models are neural network models trained on large amounts of text data in an unsupervised manner. This pre-training process allows the models to learn general linguistic knowledge from the data, including semantics, syntax, and relationships between words. PLMs can then be fine-tuned on downstream supervised tasks, such as text classification, question answering, text generation and machine translation. Well-known PLMs include BERT, GPT-2, and RoBERTa.

PLMs have shown promising results in Machine Translation. For instance, BART (Bidirectional and Auto-Regressive Transformer) is a PLM that have been used for MT and shown to improve the performance of MT systems [18]. Chronopoulou et al. [19] used a language model pre-trained on two languages with large monolingual data to initialize an unsupervised neural machine translation system, which yielded state-of-the-art results. Edunov et al. [20] have examined different strategies to integrate pre-trained representations into sequence-to-sequence models and applied it to neural machine translation. PLMs have also been used for low-resource machine translation [21], sign language translation [22], and code-mixed Hinglish-to-English machine translation [23]. However, the successful construction of such models often requires large amounts of data and computational resources [24].

Table 2 summarizes the key gaps in using PLMs for machine translation, including the lack of models tailored for particular language pairs, such as English–Arabic, insufficient data and computing access for low-resource settings, and challenges in scaling cross-lingual transfer to many languages. Our research seeks to assess the value of PLMs, even with limited resources, and evaluate their quality.

Table 2 Summary of key studies utilizing PLMs for machine translation

Approach	Key works	Findings	Limitations
PLM for MT	BART model [18] improves MT performance Chronopoulou et al. [19]: pretrained LM to initialize unsupervised NMT, SOTA results	PLMs capture linguistic knowledge useful for MT Can improve supervised and unsupervised MT	Require large data and compute resources [24]
Low-resource MT	PLMs used successfully for low-resource MT [21]	Help mitigate data scarcity challenges	Still limited by small data size
Multilingual MT	Edunov et al. [20]: integrate multilingual pretrained representations into MT models	Leverage cross-lingual transfer learning	Difficult to scale to many languages

2.3 Error Analysis in Machine Translation

Error analysis in machine translation refers to the process of identifying, categorizing, and understanding errors made by MT systems in translating text from one language to another. Despite significant improvements in translation algorithms and the application of artificial intelligence, MT systems are not error-free. Therefore, error analysis in machine translation serves as an essential process for diagnosing and refining models to improve the quality of translations and enhance comprehension.

MT systems can generate various types of errors owing to their language complexity. A study conducted by IBM highlighted the common errors observed in translation outputs when translating Russian into English. These include errors such as transliterated words, multiple meanings and ambiguities, word order rearrangements, and miscellaneous insertions and corrections [25].

Several error classification schemes have been proposed [26]. Popović [26] provided an overview of manual and automatic approaches to error classification and analysis of MT. Manual classification allows more error categories, but suffers from cost, time demands, and low annotator consistency. Automatic tools are faster and cheaper but are limited in detail and accuracy. Common machine translation error types reported by Popović include: lexical errors such as incorrect word choices or mistranslated terminology; morphological errors in inflection, derivation, and word composition; syntactic errors in word order at the word, phrase, and sentence levels; semantic errors where the meaning is changed through incorrect disambiguation or mistranslation of multi-word expressions; orthographic errors in spelling, punctuation, and capitalization; omission errors where words or phrases are missing compared to the source; addition errors where extra words or phrases are added; reordering errors where the word, phrase, or clause order differs from the source; and segments with too many errors to classify individually. Popović noted that lexical, morphological, and reordering errors are especially problematic for statistical machine translation systems. The distribution of these error types provides an "error profile" that gives insight into the performance of machine translation systems.

Similarly, Chatzikoumi [27] provided a comprehensive review of methods for evaluating machine translation quality, including both automated metrics, which compare MT output to reference translations and provide advantages such as speed and low cost but lack nuance and diagnostic feedback, and human evaluation techniques such as direct assessment, ranking, error analysis, and post-editing, which allow for more nuanced judgments but are slower, more costly, and subjective. The paper also discusses numerous error types, including mistranslations conveying incorrect or ambiguous meaning, additions of extra words without basis in the source, omissions of omitted words and phrases, incorrect translation of words that should remain unchanged, morphology errors like incorrect inflection, syntax errors with word order at the word, phrase, or sentence level, semantic errors with multiword expressions, collocations, word sense disambiguation, orthography errors in spelling, punctuation, capitalization, and fluency issues causing ungrammatical, inconsistent, or unreadable output. The paper notes that the typology and granularity of error analysis depend on the specific goals of the evaluation, whether improving a particular MT system or general quality assessment. Overall, the wide range of errors highlighted illustrates the challenges faced in machine translation and the need for rigorous, multifaceted evaluation techniques.

In summary, Table 3 provides an overview of the MT error analysis types and evaluation approaches. It is apparent that previous papers emphasize the importance but difficulty of rigorously evaluating MT quality and analyzing translation errors. Neural MT shows promise but still produces critical errors that require human evaluation. Overall, combining automated metrics and human judgment, particularly for critical semantic errors, provides the most comprehensive MT assessment.

3 Methodology

This section presents the dataset used, the PLMs, and the evaluation metrics.

Table 3 Overview of MT error analysis types and evaluation approaches

Category	Overview
Error types	Lexical, morphological, syntactic, semantic, orthographic, omission/addition, reordering, mistranslation, etc. [26, 27]
Manual analysis	Allows more error categories, but costly, slow, inconsistent [26]
Automatic analysis	Faster, cheaper, but less detailed and accurate [26]
Evaluation methods	Automatic metrics: Fast but lack nuance Human evaluation: More nuanced but slower/costly [27]

3.1 Dataset

We used data from the AEPC corpus [28] constructed to fill the gap in available Arabic-English corpora to support translation and language learning. The corpus consists of a 10-million-word Arabic-English parallel corpus crossing diverse text genres, including: social, biographical, literary, administrative, medical, legal, religious, and scientific texts. The text was manually translated, segmented into sentences, aligned, and verified for its accuracy. For this research, we chose the following genres: Psychology, Political, Medical and Scientific domains with 140, 114, 186, and 102 parallel English–Arabic sentences, respectively. The selection was based on their diverse and specialized vocabulary, differences in syntactic structures, and their potential for enhancing global information accessibility.

3.2 PLM Selection

In this study, we chose five PLMs to serve as subjects of experimentation, with a specific focus on English-to-Arabic translation. The models are divided into three open-source models: Helsinki, Marefa, Facebook, and two closed-source models: GPT-3.5-turbo and GPT-4. We also used Google Translator as a baseline model.

The Marefa model [29] was designed to cater to English-to-Arabic translation tasks. It distinguishes itself by incorporating additional Arabic characters, such as "پ" and "گ", thereby enhancing translation accuracy and preserving the fidelity of the original content.

On the other hand, the Helsinki model [30] also offers English-to-Arabic translation capabilities. It adopts a comprehensive approach to language translation by employing a diverse array of linguistic features and techniques to achieve proficient results.

The Facebook "mBART-50" model [31] represents a multilingual sequence-to-sequence model that has undergone extensive pre-training. Its introduction was accompanied by a seminal research paper titled "Beyond English-Centric Multilingual Machine Translation" [32].

GPT-3.5-turbo and GPT-4 were developed by OpenAI [33]. GPT-3.5-turbo is an extension of GPT-3 that improves its performance and efficiency in natural language understanding and generation tasks, especially for dialogue applications. GPT-4 is a multimodal model that can process both image and text inputs and generate text outputs, thereby demonstrating human-level capabilities on various professional and academic benchmarks. Both models were accessed using OpenAI API.

3.3 Evaluation Metrics

There are several methods for evaluating the performance of automatic machine translation systems, including [34]:

1. *Human evaluation* involves human judges assessing the quality of the machine translation output. The two main early methods used were ALPAC and DARPA [21]. ALPAC focuses on intelligibility (translation understandable) and fidelity (how much original information is retained). DARPA examines adequacy (how much information is conveyed), fluency (is the output grammatical), and informativeness (does it provide information about the system's abilities). However, new advanced methods, including the Multidimensional Quality Metrics (MQM) framework [35], which offers a versatile and effective evaluation framework for assessing MT quality and transcending language barriers, have gained traction among researchers owing to its adaptability and merits.
2. *Automatic evaluation* allows for faster evaluation by comparing MT outputs to human references. Some MT metrics include ChrF, which focuses on character n-grams rather than words to better handle morphologically rich languages. BLEU counts the matching n-grams between the MT and references. METEOR explicitly matches words and considers their recall and precision. It also matches the stems and synonyms. ROUGE performs comparisons based on the longest common subsequence or skip bigrams. Other more comprehensive measures include the COMET and BERTscore which are discussed next.

4 Results and Discussion

In this section, the evaluation results of the machine translation output using two approaches are presented: automatic evaluation and human evaluation.

4.1 Automatic Evaluation

chrF (character F-score) [36] This metric measures the similarity between the reference translation and the candidate translation at the character level. It calculates the F-score, which combines precision and recall, to measure overall similarity.

$$\text{chr F}\beta = (1+\beta^2)\frac{\text{CHRP} \times \text{CHRR}}{\beta^2 \times \text{CHRP} + \text{CHRR}}$$

In chrF, CHRP and CHRR denote the average precision and recall of character n-grams across all n-grams, where CHRP measures the match percentage of n-grams in the hypothesis to the reference, and CHRR measures the match in the reference to the hypothesis. The β parameter weights recall β times more than precision, with $\beta = 1$ indicating equal importance for both.

BERTScore [37] This metric leverages contextual embeddings generated by BERT, a powerful language model, to evaluate the quality of a candidate translation. It compares the candidate translation with the reference translation and assigns a similarity score.

$$\text{BERTScore} = \frac{1}{|C|}\sum_{c\in C}\max_{r\in R}\cos(e_c, e_r)$$

where |C| is the number of tokens in the candidate translation, C is the set of token embeddings in the candidate translation, R is the set of token embeddings in the reference translation, e_c and e_r are the embeddings of tokens c and r, and cos is the cosine similarity.

COMET (Cross-lingual Optimized Metric for Evaluation of Translation) [38] This is a comprehensive metric that considers different aspects of translation quality, including adequacy and fluency. It combines various evaluation dimensions to generate an overall score. It uses a neural network model, typically involving embedding layers, a transformer-based encoder, and a scoring function to evaluate translation quality. This process involves converting input sentences (source, reference, and hypothesis) into vector representations, processing them using the model to capture contextual relationships, and then outputting a quality score. The exact computation depends on the model architecture and parameters, which are trained on datasets with human translation judgments, making it a complex and dynamic evaluation method without a simple, fixed equation.

These metrics were chosen for their ability to comprehensively evaluate translation quality, capture nuances at the character level, semantic similarity via contextual embeddings, and holistic quality measurements, including fluency and adequacy.

Table 4 shows the evaluation results of the six different machine translation models—Google Translate, Helsinki, Marefa, and Facebook; GPT-3.5-Turbo; and GPT-4–across four domains: Psychology, Political, Medical, and Scientific, using three automatic evaluation metrics: ChrF, BERTscore, and COMET.

The results revealed notable differences in performance among the various machine translation models when translating from English to Arabic. GPT-4 and gpt-3.5-turbo emerged as the top performers, demonstrating superior capabilities across all evaluation metrics. This was particularly evident in their handling of semantic and contextual nuances, as reflected by their high BERTscore and COMET scores. However, the Helsinki model exhibited limitations, with lower average scores in all domains, suggesting potential gaps in its translation algorithms or training data for this language pair. Google Translator showed robust performance, especially in semantic and contextual understanding, which is critical for effectively translating nuanced texts.

In terms of domain-specific performance, translations within the scientific domain achieved the highest average ChrF scores, indicating a better alignment at the character level. This could be attributed to the technical and less-idiomatic nature of scientific texts. The Medical domain translations showed superior accuracy in context and semantics, as evidenced by the highest BERTscore and COMET scores. This suggests that these models are particularly effective for texts with standardized terminology and structures. However, the variability in model performance in the Psychology and Political domains implies that these areas possess the linguistic and contextual complexities that current machine translation models find challenging.

Table 4 MT automatic evaluation results

Domain	Model	ChrF	BERTscore	COMET
Psychology	Google Translator	0.518	**0.857**	0.845
	Helsinki	0.156	0.825	0.764
	Marefa	0.478	0.825	0.841
	Facebook	0.372	0.825	0.841
	gpt-3.5-Turbo	0.608	0.8418	**0.851**
	GPT-4	**0.610**	0.8474	0.850
Political	Google Translator	0.493	**0.846**	0.844
	Helsinki	0.081	0.825	0.815
	Marfa	0.476	0.823	0.816
	Facebook	0.488	0.833	0.826
	gpt-3.5-Turbo	0.618	0.8342	**0.858**
	GPT-4	**0.621**	0.839	0.857
Medical	Google Translator	0.471	**0.847**	**0.867**
	Helsinki	0.410	0.821	0.845
	Marefa	0.403	0.822	0.844
	Facebook	0.426	0.836	0.853
	gpt-3.5-Turbo	0.596	0.8334	0.857
	GPT-4	**0.603**	0.8426	0.865
Scientific	Google Translator	0.556	**0.861**	**0.835**
	Helsinki	0.458	0.8235	0.8
	Marefa	0.439	0.8246	0.791
	Facebook	0.442	0.8248	0.804
	gpt-3.5-Turbo	0.615	0.8289	0.805
	GPT-4	**0.625**	0.8388	0.808

Bold font indicates best result obtained

Our findings are significant in the broader context of machine translation, particularly for the English-to-Arabic language pair, which is often challenged by structural and contextual differences. The high performance of advanced models such as GPT-4 and gpt-3.5-turbo marks a significant step forward in overcoming language barriers, thereby showcasing the potential of AI in this field. This study also highlights the importance of considering domain-specific nuances in machine translation, underlining the need for tailored approaches and enhancements in translation models for different domains.

When compared with the existing literature, our study aligns with previous research on the varying efficacy of translation models across different text types but extends this understanding to a more nuanced analysis of domain-specific performance [39]. The success of models such as GPT-4 and GPT-3.5-turbo corroborates recent studies on the effectiveness of large-language models in translation tasks. Conversely, the observed limitations in models, such as Helsinki, echo broader research challenges, particularly for complex linguistic structures or lesser-resourced languages. Our study contributes to this discourse by identifying the specific domains in which these challenges are more pronounced.

4.2 Human Evaluation

As mentioned before, we chose to employ the Multidimensional Quality Metrics (MQM) framework for conducting the manual evaluation task, focusing on error analysis. The MQM framework functions as a valuable instrument for defining and building personalized translation quality metrics. It provides a flexible set of quality issue categories and a means to utilize them, resulting in the respective quality scores.

Thus, to assess the correlation between automatic metrics and human judgment in the context of MT, a representative sample of 106 sentences was selected from the four domains. The human translations of these sentences were compared with the translations generated by the six MT systems mentioned earlier. Using human translation as a reference point, we can conduct a thorough evaluation of the MT system's *accuracy* and *fluency* in capturing the intended meaning and linguistic quality of the source text.

Table 5 lists the relevant error analysis categories used in this study. We examined the sentences and examined them sentence by sentence and word by word to detect any translation errors.

Table 6 presents the error rates for the various machine translation systems. Each system was used to translate a set of sentences and the error rate represented the percentage of incorrect translations. Lower percentages indicate better performance in terms of accuracy. From this analysis, we can see that Google translate had the lowest error rate, indicating that it performed the best among the evaluated systems. On the other hand, Facebook translations exhibited the highest error rate.

It is noteworthy that these error rates represent the overall translation quality of each system. Further examination of the error distribution reveals that Addition, Omission, Mistranslation, Untranslated and Grammar categories were the primary sources of errors across the systems, as shown in Fig. 1. In summary, the data suggest that Google Translate outperforms the other MT systems, whereas Helsinki exhibits the highest error rate, indicating potential areas for improvement in these systems.

Based on the analysis of translation accuracy and fluency across six different machine translation (MT) systems, several insights have emerged. Google demonstrated the most robust performance, with the lowest total number of errors,

Table 6 Error rates of different machine translation systems

Translation system	Error rate (%)
Google	32.5
Helsinki	67.5
Marefa	67.5
Facebook	72.5
GPT-3.5-Turbo	55.0
GPT4	45.0

Table 5 The set of employed error categories

Main category	Subcategory	Definition
Accuracy	Addition	The target includes information not present in the source, for example, adding a date that does not exist in the source text to the translation
	Omission	Content is missing from the translation that is present in the source, for instance, deleting the negation in the translation
	Mistranslation	The target content does not accurately represent the source content
	Untranslated	Content that should have been translated has been left untranslated
Fluency	Spelling	A word is misspelled
	Grammar	Issues related to the grammar or syntax of the text such as function words, word order, agreement, tense, and parts of speech

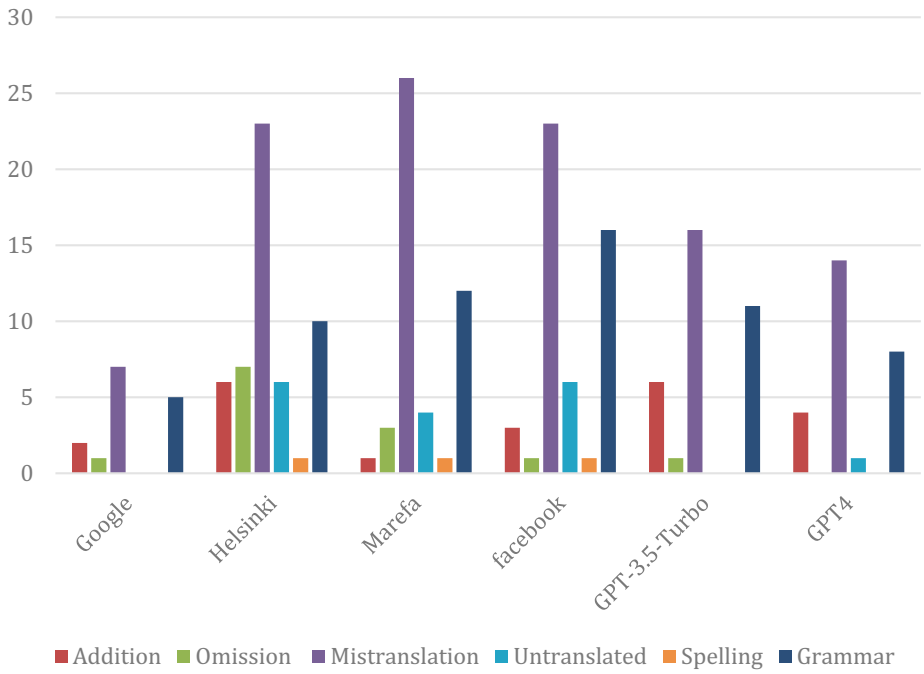

Fig. 1 Translation error counts and distribution of major categories in machine translation systems

amounting to only 15 across the evaluated categories. In contrast, Helsinki exhibited the highest error count with 53 mistakes, suggesting significant challenges in its translation mechanism. Marefa and Facebook also show a relatively high error rate, with 47 and 50 errors respectively, indicating areas for improvement, particularly in handling mistranslations, which emerge as the most common error type across all platforms.

Interestingly, the newer GPT models, GPT-3.5-Turbo and GPT4, exhibited intermediate performance. GPT-3.5-Turbo registers 34 errors, while GPT4 accounts for 27, positioning them better than Helsinki and Facebook, but still trailing behind Google's superior accuracy. This pattern underscores Google's continued leadership in the MT domain despite the advancements and introduction of newer technologies such as GPT models.

Mistranslation stands out as a universal challenge for all assessed systems, highlighting it as a primary difficulty in machine translation. This consistent issue across different platforms points to the inherent complexities of achieving accurate and contextually relevant translations. Furthermore, the specific breakdown of errors such as Grammar, Omission, and Untranslated sections offers a more granular view of each system's strengths and weaknesses, providing valuable insights for future improvements in machine translation technologies. This comprehensive analysis sheds light on the current state of MT systems, revealing both their achievements and limitations in dealing with language translation tasks.

Table 7 presents a comprehensive summary of the error rates of different machine translation systems across various disciplines. This table provides a clear comparison of the performance of each translation system across different fields, highlighting the variability and specific challenges in each domain. We can observe distinct patterns in translation quality across different text genres and translation systems. The following is a detailed analysis.

Table 7 Error rates of different machine translation systems across various disciplines

Discipline	Google error rate (%)	Helsinki error rate (%)	Marefa error rate (%)	Facebook error rate (%)	GPT-3.5-Turbo error rate (%)	GPT4 error rate (%)
Medical	30.0	20.0	30.0	70.0	30.0	10.0
Psychological	40.0	70.0	80.0	90.0	70.0	50.0
Political	20.0	80.0	60.0	50.0	40.0	30.0
Scientific	40.0	100.0	100.0	80.0	80.0	90.0

1. Medical Domain
   - In the medical field, where accuracy is critical, we observed varied performances across different translation systems. Helsinki and GPT4 showed promising results, indicating their potential utility in medical translation. However, the relatively higher error rates in other systems, particularly Facebook, highlight the need for caution, and possibly human oversight, to ensure precision and reliability.
2. Psychological Domain
   - Psychological texts, characterized by complex and nuanced languages, present a significant challenge for all translation systems. The higher error rates across the board suggest that machine translations in this field require extensive review and editing. These findings emphasize the importance of understanding the limitations of machine translation in handling the subtleties and specificities of psychological terminology.
3. Political Domain
   - Political texts, with their nuanced language and context sensitivity, show a wide range of performance among different systems. Some systems demonstrate a better grasp of political language, whereas others exhibit notable difficulties. This variability underscores the importance of selecting the correct translation tool for political content and the need for careful review, especially when the text is intended for sensitive or critical use.
4. Scientific Domain
   - The scientific field, known for its technical jargon and necessity for precise language, poses the greatest challenge for machine translation systems. The high error rates across almost all systems indicate that while machine translations can provide a starting point, they often fail to accurately capture the technical nuances of scientific texts. This finding reinforces the need for expert review and suggests that relying solely on machine translation in this domain may be inadequate.

In conclusion, the analysis highlights the critical role of human oversight, particularly in fields where accuracy and context are paramount. Machine translations, while beneficial as a starting point, should be complemented with human expertise to ensure fidelity and accuracy, particularly in scientific, psychological, and political texts. In the next section, we provide a detailed error analysis of the results to understand the root causes of such errors.

4.3 Error Analysis

To gain a better understanding of the types of errors committed by the above machine translation systems, some examples of classification errors are presented next.

(1) Analysis of Fluency Errors

Fluency errors address issues related to the grammar or syntax of the text, such as function words, word order, agreement, tense, and parts of speech.

The examples in Table 8 highlight several key error categories in machine translation. A notable issue observed in Google's translate is the omission of the Arabic article "ال," reflecting a common grammatical error in translating functional morphemes. This omission affects the grammatical integrity of the Arabic sentences.

Facebook's translation exhibits an error in subject-verb agreement, particularly with respect to gender, underscoring the challenge of maintaining gender agreement, which is crucial in Arabic grammar. Additionally, Facebook's translation demonstrates a problem with word order, an error that can significantly impact the clarity and coherence of translated text in Arabic. Another type of error observed in Facebook's translation is the incorrect translation of a singular subject into a plural form, leading to a semantic discrepancy that can alter the intended meaning of the sentence. Furthermore, Facebook also shows a tendency to use the wrong form of a word, an error that can result in misunderstanding or a change in the sentence's intended meaning.

Each of these errors underlines the specific challenges faced in machine translation, particularly when dealing with the complexities of Arabic language structure, including grammar, word order, and agreement.

(2) Analysis of Accuracy Errors

In the domain of machine translation accuracy, various error categories have been identified for in-depth analysis, including addition, omission, mistranslation, and issues with untranslated words, as shown in Table 9. This analysis aimed to uncover the intricacies and challenges inherent in translating English and Arabic.

For instance, in the realm of addition errors, we observed a case of medical translation. The original English sentence, "What do we mean by body language?", when translated by Facebook, becomes "ما الذي نعنيه باستعمال لغة الجسد؟". Here, the phrase "باستعمال" ("by using") is added, subtly shifting the meaning to "What do we mean by using body language?" This addition, while not drastically altering the message, is unnecessary for conveying the fundamental essence of the sentence, as demonstrated by a human translator's rendition: "ماذا نقصد بلغة الجسد ؟".

Google's translation of a psychological sentence further illustrates the addition errors. The sentence "These elements are referred to as 'para-linguistic cues'" is translated to "يشار إلى هذه العناصر باسم 'الإشارات شبه اللغوية'." Here, the insertion of "باسم" and the definite article "ال" before "إشارات شبه اللغوية" ("para-linguistic cues") implies a

Table 8 Examples of MT fluency errors and their issues

Source/English	Arabic/human translation	Issue
I never thought of these as bad qualities	ولم أنظر أبدًا لهذه الصفات على أنها صفات سيئة	Grammar/omitting functional morphemes (article ال)
	Google	
	لم افكر ابدا في هذه صفات سيئة	
Only a few weeks after that historic night of December 3, 1967 a French company offered me $50,000 for the surgical gloves I had worn during the operation	Arabic/human translation	Violating subject-verb agreement (masculine and feminine) Wrong word order
	فبعد أسابيع من تلك الليلة التاريخية الثالث من شهر ديسمبر من عام 1967، عرضت علي شركة فرنسية مبلغ 50 ألف دولار مقابل القفازات الطبية التي ارتديتها في العملية	
	Facebook	
	بعد أسابيع قليلة فقط من تلك الليلة التاريخية من 3 ديسمبر 1967 قدم لي شركة فرنسية 50 ألف دولار لقفاز الجراحة التي كنت أرتديها خلال العملية	
	GPT-3.5-Turbo	
	بضعة أسابيع فقط بعد تلك الليلة التاريخية في الثالث من ديسمبر 1967، عرضت لي شركة فرنسية 50,000 دولار مقابل القفازات الجراحية التي كنت قد ارتديتها أثناء العملية	
This wondrous organ has only one task to perform in our bodies—it pumps blood	Arabic/ Human Translation	Wrong translation of singular subject أعضاء → عضو
	وهذا العضو المثير للعجب لديه وظيفة واحدة يؤديها في أجسامنا: ضخ الدم	
	Facebook	
	هذه الأعضاء المذهلة لديها مهمة واحدة فقط لتقوم بها في أجسادنا—أنها تضخ الدم	
Expressions—the arrangement of the face and eye movements that convey a great deal of the meaning of a communication	Arabic/human translation	Using wrong form of the word
	العبارات—هي ترتيب حركات الوجه والعين التي تنقل مقدارًا كبيرًا من المعاني للاتصال	
	Facebook	
	التعبيرات- ترتيب حركة الوجه والعين التي تنقل الكثير من معنى التواصل	

Table 9 Examples of MT accuracy errors and their issues

Accuracy error	Source/English	Arabic/human translation	Issue
Addition	What do we mean by body language?	"ماذا نقصد بلغة الجسد ؟" Facebook ما الذي نعنيه باستعمال لغة الجسد؟	Addition of "باستعمال" ("by using") alters the meaning
	These elements are referred to as 'paralinguistic cues'	Arabic/human translation وتعتبر هذه العناصر "إشارات شبه لفظية Google يشار إلى هذه العناصر باسم "الإشارات شبه اللغوية"	Addition of "باسم" and "ال" changes the specific reference and meaning
	In a crisis, cash becomes king	Arabic/human translation في الأزمات يصبح المال هو الحاكم GPT-3.5-Turbo في حالة الأزمات، يصبح النقد هو الأهم	Inclusion of "حالة" ("case/situation") introduces an unnecessary element
Omission	What we tend to do is interpret what people are saying with reference to their body language	Arabic/human translation فنحن نميل إلى تفسير مايقوله الناس بالرجوع إلى لغة جسدهم Helsinki ما نميل إلى فعله هو تفسير ما يقوله الناس بالإشارة إلى لغتهم	Omission of "body" before "language," altering the intended meaning
	I simply have not been able to come up with anything that would have made a difference	Arabic/human translation ببساطة لم أكن قادرًا على التوصل إلى أي شيء كان من الممكن أن يشكل فارقاً Marefa لقد تمكنت ببساطة من التوصل إلى أي شيء كان من شأنه أن يحدث فرقا	Omission of "not," changing the meaning to the opposite of what is intended
Mistranslation	It seems incredible, but it contains more than 300 references to the heart	Arabic/human translation وكان من المذهل أن أجد أنه يحتوي على أكثر من 300 إشارة للقلب Facebook يبدو الأمر مذهلاً، لكنه يحتوي على أكثر من 300 مرجعاً للقلب	The target content does not accurately represent the source content
Untranslated word	How NLP contributes to understanding body language	Arabic/ Human Translation كيف تساهم البرمجة اللغوية العصبية في فهم لغة الجسد Facebook كيف تساهم الـ NLP في فهم لغة الجسد	Untranslated word

specificity and recognition in Arabic that the original English sentence does not suggest.

Similarly, GPT-3.5-Turbo's translation of "In a crisis, cash becomes king" to "في حالة الأزمات، يصبح النقد هو الأهم" introduces the word "حالة" ("case" or "situation") unnecessarily, as it was absent in the original English phrase.

On the omission front, Helsinki's translation of "What we tend to do is interpret what people are saying with reference to their body language" to "ما نميل إلى فعله هو تفسير ما يقوله الناس بالإشارة إلى لغتهم" misses the critical word "body" before "language." This omission significantly alters the intended meaning and reduces the specificity of interpreting verbal communication alongside body language.

Moreover, Marefa's translation of "I simply have not been able to come up with anything that would have made a difference" as "لقد تمكنت ببساطة من التوصل إلى أي شيء كان من شأنه أن يحدث فرقاً" omits the crucial word "not." This negates the intended meaning and falsely suggests that the speaker was able to come up with a significant idea, contrary to the source's expression of inability.

According to the issue of mistranslation, the original English sentence expresses amazement with a straightforward structure. The translation provided by Facebook does not accurately represent source content in several ways. Firstly, the phrase "يبدو الأمر مذهلاً" ("The matter seems amazing") changes the nuance of the sentence. The original phrase "It

seems incredible" conveys a sense of disbelief or astonishment at the quantity of references. However, the Facebook translation shifts this to a more general sense of amazement, losing the subtlety of incredulity present in the original version. Secondly, the use of "مرجعاً" (references) in the translation may not capture the intended nuance of the English word "references." In English, "references to the heart" can imply various types of mentions or allusions, not just formal citations or sources. The Arabic translation could be interpreted as more formal or academic, which might not perfectly align with the original English expression.

In the untranslated word issue, the acronym "NLP" is left untranslated. This untranslated word presents a significant issue in terms of accessibility and understanding for Arabic-speaking audiences, who may not be familiar with the English acronym. "NLP" stands for "Neuro-Linguistic Programming," a concept that would typically be translated into Arabic as "البرمجة اللغوية العصبية." Leaving "NLP" untranslated can potentially lead to confusion or misinterpretation among readers who are not accustomed to English acronyms or who may not recognize the acronym in the context of the Arabic language. It is crucial in translation, especially when dealing with technical or specialized terms, to ensure that such terms are appropriately translated to maintain the clarity and comprehensibility of the content for the target audience.

Each of these examples underscores the nuanced challenges in machine translation, especially in dealing with Arabic's complex grammar, word order, and agreement. These findings highlight the need for deeper understanding and more sophisticated approaches to overcome these challenges and ensure more accurate and effective translations.

5 Conclusion and Future Work

In conclusion, this study investigated the performance of state-of-the-art pretrained language models (PLMs) on English-to-Arabic machine translation across diverse text genres. Our analysis revealed that advanced models, such as GPT-4 and GPT-3.5-turbo, demonstrate superior translation capabilities based on automatic metrics, such as chrF, BERTscore, and COMET. However, all systems still face challenges in accurately conveying complex grammar, vocabulary, and meaning when handling texts in domains such as psychology and political science.

Through manual evaluation using the MQM framework, we identified lexical, morphological, syntactic, semantic, and fluency issues that pose difficulties to PLMs. Key error patterns include omissions or incorrect translations of functional words, lack of subject-verb gender agreement, word order rearrangements, mistranslations from ambiguity, and spelling/grammar mistakes that disrupt fluency. Google Translate is currently the most robust, whereas the other systems show higher domain-specific variability.

In summary, our study demonstrates the promise of pretrained models in advancing neural machine translation quality but also highlights persistent limitations in terms of Arabic grammar, terminology, and contextual nuances. As English–Arabic machine translation holds significance for global communication and cooperation, our analysis offers valuable insights into pressing challenges and future priorities for the field. Specifically, our work highlights the need for continued research on tailored architectures, multilingual representations, contextual encoding, and specialized model training to further enhance English-to-Arabic translation performance across diverse real-world texts.

Acknowledgements We acknowledge the use of ChatGPT, an AI chatbot developed by OpenAI, for generating some of the summaries in this article. ChatGPT was used to supplement our own writing and analysis, and not to replace them. We verified the accuracy and relevance of the AI-generated text before incorporating it into our manuscript.

Author Contributions HK: methodological framework, experiments, evaluation, writing preliminary manuscript draft; revising final manuscript; supervision; KK: experimental result analysis, reviewing and editing manuscript. HH experiments; All authors approved the final version of the manuscript.

Funding This study was funded by the Literature, Publishing and Translation Commission, Ministry of Culture, Kingdom of Saudi Arabia under [73/2022] as part of the Arabic Observatory of Translation.

Availability of Data and Materials No underlying data were collected or produced for this study.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Ethical approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication The authors hereby grant full consent for the publication of the manuscript in the HCIN journal.

References

1. Zaugg IA, Hossain A, Molloy B. Digitally-disadvantaged languages. Internet Policy Rev J Internet Regul. 2022;11(2):1–11.
2. Patil A, Joshi I, Kadam D. PICT@WAT 2022: neural machine translation systems for indic languages. In: Proceedings of the 9th workshop on Asian Translation, Gyeongju, Republic of Korea: international conference on computational linguistics. 2022. pp. 106–110. https://aclanthology.org/2022.wat-1.13. Accessed 20 Dec 2023.
3. Chen K, Wang R, Utiyama M, Sumita E. Integrating prior translation knowledge into neural machine translation. IEEEACM Trans Audio Speech Lang Process. 2022;30:330–9. https://doi.org/10.1109/TASLP.2021.3138714.
4. Akan MF, Karim MR, Chowdhury AMK. An analysis of Arabic–English translation: problems and prospects. Adv Lang Lit Stud. 2019;10(1):58–65. https://doi.org/10.7575/aiac.alls.v.10n.1p.58.
5. Mamoori MMA, Tarish AH, Hasani SA. Difficulties of translation and evaluative idioms in English and Arabic. Int J Health Sci. 2022. https://doi.org/10.53730/ijhs.v6nS5.10039.
6. Mars M. From word embeddings to pre-trained language models: a state-of-the-art walkthrough. Appl Sci. 2022;12(17):art no. 17. https://doi.org/10.3390/app12178805.
7. Zakraoui J, Saleh M, Al-Maadeed S, AlJa'am JM. Evaluation of Arabic to English machine translation systems. In: 2020 11th International conference on information and communication systems (ICICS). 2020. pp. 185–190. https://doi.org/10.1109/ICICS49469.2020.239518.
8. Bar-Hillel Y. The Present status of automatic translation of languages. In: Alt FL, editors. Advances in computers, vol. 1. Elsevier; 1960. pp. 91–163. https://doi.org/10.1016/S0065-2458(08)60607-5.
9. Ameur MSH, Meziane F, Guessoum A. Arabic machine translation: a survey of the latest trends and challenges. Comput Sci Rev. 2020;38: 100305. https://doi.org/10.1016/j.cosrev.2020.100305.
10. Zakraoui J, Saleh M, Al-Maadeed S, Alja'am JM. Arabic machine translation: a survey with challenges and future directions. IEEE Access. 2021;9:161445–68. https://doi.org/10.1109/ACCESS.2021.3132488.
11. Farhat A, Al-Taani AT. A rule-based English to Arabic machine translation approach. In: Presented at the international Arab conference on information technology (ACIT'2015). 2015. https://www.semanticscholar.org/paper/A-Rule-based-English-to-Arabic-Machine-Translation-Farhat-Al-Taani/4e7f555a0221eb7f980c597b15bdb8f6a1089e7f. Accessed 16 Jul 2023.
12. Fadiel Alawneh M, Sembok TM, Mohd M. Grammar-based and example-based techniques in machine translation from English to Arabic. In: 2013 5th international conference on information and communication technology for the Muslim World (ICT4M). 2013. pp. 1–6. https://doi.org/10.1109/ICT4M.2013.6518910.
13. Al-Rukban A, Saudagar AKJ. Evaluation of English to Arabic machine translation systems using BLEU and GTM. In: Proceedings of the 2017 9th international conference on education technology and computers. ACM; 2017.
14. Akeel M, Mishra R. ANN and rule based method for english to arabic machine translation. Int Arab J Inf Technol. 2014;11(4):396–405.
15. Aljohany DA, Al-Barhamtoshy HM, Abukhodair FA. Arabic machine translation (ArMT) based on LSTM with attention mechanism architecture. In: 2022 20th International conference on language engineering (ESOLEC). 2022. pp. 78–83. https://doi.org/10.1109/ESOLEC54569.2022.10009530.
16. Aref M, Al-Mulhem M, Al-Muhtaseb H. English to Arabic machine translation: a critical review and suggestions for development. King Fahd Univ. Pet. Miner. Dhahran Saudi Arab. 1992.
17. Nagoudi EMB, Elmadany A, Abdul-Mageed M. TURJUMAN: a public toolkit for neural Arabic machine translation. In: Proceedings of the 5th workshop on open-source Arabic corpora and processing tools with shared tasks on Qur'an QA and fine-grained hate speech detection. Marseille, France: European Language Resources Association; 2022. pp. 1–11. https://aclanthology.org/2022.osact-1.1. Accessed 16 Jul 2023.
18. Lewis M, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2020. pp. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703.
19. Chronopoulou A. Stojanovski D, Fraser A. Reusing a pretrained language model on languages with limited corpora for unsupervised NMT. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics; 2020. pp. 2703–2711. https://doi.org/10.18653/v1/2020.emnlp-main.214.
20. Edunov S, Baevski A, Auli M. Pre-trained language model representations for language generation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. pp. 4052–4059. https://doi.org/10.18653/v1/N19-1409.
21. Zheng F, Reid M, Marrese-Taylor E, Matsuo Y. Low-resource machine translation using cross-lingual language model pretraining. In: Proceedings of the first workshop on natural language processing for indigenous languages of the Americas. Association for Computational Linguistics; 2021. pp. 234–240. https://doi.org/10.18653/v1/2021.americasnlp-1.26.
22. De Coster M, Dambre J. Leveraging frozen pretrained written language models for neural sign language translation. Information. 2022;13(Art. no. 5):5. https://doi.org/10.3390/info13050220.
23. Agarwal V, Rao P, Jayagopi DB (2023) Hinglish to English machine translation using multilingual transformers. In: Proceedings of the student research workshop associated with RANLP 2021, INCOMA Ltd., Sep. 2021, pp. 16–21. https://aclanthology.org/2021.ranlp-srw.3. Accessed 16 Jul 2023.
24. Jude Ogundepo O, Oladipo A, Adeyemi M, Ogueji K, Lin J. AfriTeVA: Extending? Small data? Pretraining approaches to sequence-to-sequence models. In: Proceedings of the third workshop on deep learning for low-resource natural language processing, hybrid. Association for Computational Linguistics; 2022. pp. 126–135. https://doi.org/10.18653/v1/2022.deeplo-1.14.
25. Hutchins WJ. Machine translation: a brief history. In: Koerner EFK, Asher RE, editors. Concise history of the language sciences. Amsterdam: Pergamon; 1995. p. 431–45. https://doi.org/10.1016/B978-0-08-042580-1.50066-0.
26. Popović M. Error classification and analysis for machine translation quality assessment. In: Moorkens J, Castilho S, Gaspari F, Doherty S, editors. Translation quality assessment: from principles to practice. Machine translation: technologies and applications. Cham: Springer International Publishing; 2018. pp. 129–158. https://doi.org/10.1007/978-3-319-91241-7_7.
27. Chatzikoumi E. How to evaluate machine translation: a review of automated and human metrics. Nat Lang Eng. 2020;26(2):137–61. https://doi.org/10.1017/S1351324919000469.
28. Alotaibi H (2023) Arabic-English parallel corpus: a new resource for translation training and language teaching. Arab World Engl J AWEJ 2017;8(3). https://awej.org/arabic-english-parallel-corpus-a-new-resource-for-translation-training-and-language-teaching/. Accessed 26 Jul 2023.
29. marefa-nlp/marefa-mt-en-ar · Hugging Face. https://huggingface.co/marefa-nlp/marefa-mt-en-ar. Accessed 19 Jul 2023.

30. Helsinki-NLP/opus-mt-tc-big-ar-en · Hugging Face. https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-ar-en. Accessed 19 Jul 2023.
31. facebook/m2m100_1.2B · Hugging Face. https://huggingface.co/facebook/m2m100_1.2B. Accessed 19 Jul 2023.
32. Fan A, et al. Beyond English-centric multilingual machine translation. J Mach Learn Res. 2021;22(1):107:4839-107:4886.
33. OpenAI Platform. https://platform.openai.com. Accessed 26 Jul 2023.
34. Mondal SK, Zhang H, Kabir HMD, Ni K, Dai H-N. Machine translation and its evaluation: a study. Artif Intell Rev. 2023;56(9):10137–226. https://doi.org/10.1007/s10462-023-10423-5.
35. Lommel A. Metrics for translation quality assessment: a case for standardising error typologies. In: Moorkens J, Castilho S, Gaspari F, Doherty S, editors. Translation quality assessment: from principles to practice. Machine translation: technologies and applications. Cham: Springer International Publishing; 2018, pp. 109–127. https://doi.org/10.1007/978-3-319-91241-7_6.
36. Popović M. chrF: character n-gram F-score for automatic MT evaluation. In: Bojar O, Chatterjee R, Federmann C, Haddow B, Hokamp C, Huck M, Logacheva V, Pecina P, editors. Proceedings of the tenth workshop on statistical machine translation. Lisbon, Portugal: Association for Computational Linguistics; 2015. pp. 392–395. https://doi.org/10.18653/v1/W15-3049.
37. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. 2020. https://doi.org/10.48550/arXiv.1904.09675.
38. Rei R, Stewart C, Farinha AC, Lavie A. COMET: a neural framework for MT evaluation. 2020. https://doi.org/10.48550/arXiv.2009.09025.
39. Lyu C, Xu J, Wang L. New trends in machine translation using large language models: case examples with ChatGPT. 2023. https://doi.org/10.48550/arXiv.2305.01181.