Evaluating the Performance of Large Language Models on Arabic Lexical Ambiguities: A Comparative Study with Traditional Machine Translation Systems

Hamad Abdullah H Aldawsari[1]

[1] Department of English, College of Sciences and Humanities, Prince Sattam Bin AbdulAziz University, Wadi ad-Dawasir, Saudi Arabia

Correspondence: Hamad Abdullah H Aldawsari, Department of English, College of Sciences and Humanities, Prince Sattam Bin AbdulAziz University, Wadi ad-Dawasir, Saudi Arabia. E-mail: h.alnawwar@psau.edu.sa

Received: August 4, 2024 Accepted: October 29, 2024 Online Published: January 24, 2025

doi:10.5430/wjel.v15n3p354 URL: https://doi.org/10.5430/wjel.v15n3p354

Abstract

The rapid advancement in natural language processing (NLP) has led to the development of large language models (LLMs) with impressive capabilities in various tasks, including machine translation. However, the effectiveness of these new systems in handling linguistic complexities, such as Arabic lexical ambiguity, remains underexplored. This study investigates whether LLMs can outperform traditional machine translation (MT) systems in translating Arabic lexical ambiguities, characterized by homonyms, heteronyms, and polysemes. The evaluation involves two prominent LLMs, OpenAI's GPT and Google's Gemini, and compares their performance with traditional MT systems, Google Translate and SYSTRAN. The results indicate that GPT and Gemini offer substantial improvements in translation accuracy and intelligibility over traditional MT systems and highlight the advanced capabilities of LLMs in handling the complexities of Arabic, suggesting a significant step forward in machine translation technologies. This study highlights the potential of LLMs to overcome the limitations of traditional MT systems and provides a foundation for future research. The results contribute to the ongoing development of more effective and accurate translation systems, emphasizing the importance of adopting advanced AI technologies in the field of machine translation.

Keywords: Arabic lexical ambiguity, translation studies, Large Language Model systems, machine translation, GPT, Gemini

1. Introduction

New breakthroughs in natural language processing (NLP) have spurred the development of large language models (LLMs), which excel at a wide range of tasks, including machine translation. However, despite their potential, it is essential to evaluate these new systems' ability to handle linguistic complexities that traditional machine translation systems struggle with. One such challenge is Arabic lexical ambiguity, characterized by homonyms, heteronyms, and polysemes. While previous studies (e.g., Al-Kharabsheh & Yassin, 2017; At-tall, 2019) demonstrated the difficulties faced by popular machine translation systems like Google Translate and SYSTRAN when dealing with complex Arabic linguistic features, more recent research (e.g., Aldawsari, 2024; Boughorbel & Hawasly, 2023; Banimelhem & Amayreh, 2023) suggest potential of LLM systems in dealing with challenging translations.

The aim of this research paper is to investigate whether LLMs can outperform traditional MT systems in handling Arabic lexical ambiguity. Two prominent large language models (LLMs), OpenAI’s GPT (GPT) and Google’s Gemini (Gemini), were selected for evaluation against Google Translate and SYSTRAN. The evaluation utilized a test suite of sixteen Arabic sentences, each designed to present challenging examples of lexical ambiguity. By comparing the performance of these four systems, the study hopes to shed light on whether LLMs can overcome the limitations faced by traditional MT systems when dealing with Arabic lexical ambiguity.

Machine translation systems face significant challenges when translating languages with rich morphological structures and high lexical ambiguity, such as Arabic. Traditional MT systems, including Google Translate and SYSTRAN, have shown limitations in accurately translating ambiguous Arabic words, leading to errors and misunderstandings (Aldawsari, 2023, Al-Kharabsheh & Yassin, 2017; At-tall, 2019). This highlights the importance of assessing whether the innovative capabilities of LLM systems can overcome these challenges and deliver more accurate, contextually appropriate translations.

This study is significant for several reasons. First, it provides an empirical evaluation of the performance of LLMs in translating Arabic lexical ambiguities, contributing to the broader field of machine translation research. By identifying the strengths and weaknesses of LLMs compared to traditional MT systems, this study offers insights into the current state of translation technology and its potential for improvement. Additionally, the findings of this study will be valuable for developers and users of MT systems, providing practical recommendations for improving translation accuracy and reliability. Ultimately, this research aims to enhance the quality of Arabic-English translations, benefiting various domains such as education, business, and international communication.

To achieve the objectives of this study, the following research questions were formulated:

1. How accurately do OpenAI’s GPT and Google’s Gemini translate ambiguous Arabic words compared to Google Translate and SYSTRAN?
2. In what ways do the performances of GPT and Gemini differ from those of Google Translate and SYSTRAN when handling homonyms, heteronyms, and polysemes in Arabic?
3. What are the specific strengths and weaknesses of GPT and Gemini in dealing with Arabic lexical ambiguity, and how do they compare to traditional MT systems?

By addressing these questions, this study aims to provide a comprehensive evaluation of LLMs' capabilities in handling Arabic lexical ambiguity, offering valuable insights into the potential advancements in machine translation technology.

2. Literature Review

2.1 Machine Translation and Its Evolution

Machine Translation (MT) has undergone significant development since its inception in the mid-20th century. Initially, MT systems relied on rule-based and statistical methods to translate text from one language to another. These early approaches faced considerable limitations in accurately capturing the complexities of human language (Hutchins, 2001; Hutchins & Somers, 1992; Koehn, 2010). Despite these challenges, the field of MT has seen substantial progress with the advent of neural machine translation (NMT) systems, which utilize deep learning techniques to improve translation quality (Bahdanau, Cho, & Bengio, 2015, Zakraoui, Saleh, Al-Maadeed & Alja’am, 2021).

Google Translate and Bing Translator are among the most widely used MT systems today. Google Translate, which initially employed a purely statistical-based approach, has since transitioned to NMT, allowing it to process and translate entire sentences rather than just individual words or phrases (Turovsky, 2016). Similarly, Bing Translator has evolved to incorporate advanced machine learning algorithms, enhancing its ability to handle complex linguistic structures (Almahasees, 2021).

2.2 Challenges in Translating Arabic

Arabic poses unique challenges for MT systems, primarily due to its intricate morphology, wide variety of dialects, and frequent use of words with multiple meanings. Lexical ambiguity in Arabic, including homonyms (words with identical spelling and pronunciation but different meanings), heteronyms (words with identical spelling but different pronunciation and meanings), and polysemes (words with related but distinct meanings), makes accurate translation particularly difficult (Al-Kharabsheh & Yassin, 2017; Ameur, Meziane & Guessoum, 2020, Okpor, 2014). Research has shown that traditional MT systems struggle with these complexities, resulting in subpar translation quality for Arabic texts (Ali, 2020; Al-Jarf, 2023; Habash & Diab, 2012; Harrat, Meftouh & Smaili, 2019). A notable study by At-tall (2019) compared the performance of Google Translate and human translators in rendering colloquial Arabic expressions from speeches by the late Prime Minister Wasfi At-Tall into English. The findings highlighted the persistent challenges faced by MT systems, which produced “low quality” translations, failed to accurately convey the intended meanings of colloquial expressions and “directly followed the literal meaning” (p. 42). However, a more recent study (Aldawsari, 2024) found out that LLM tools produced significantly more accurate translations of the same data-set, highlighting a potential “advantage” of these advanced tools in overcoming the challenges of translating Arabic lexical ambiguities (p. 237).

2.3 Advantages of Large Language Models

Recent advancements in Natural Language Processing (NLP) have led to the development of Large Language Models (LLMs) such as OpenAI's GPT series and OpenAI’s GPT. These models leverage vast datasets and sophisticated algorithms to generate more accurate translations. However, studies suggest that LLMs may still fall short in handling specific linguistic features, particularly those found in Arabic (Boughorbel & Hawasly, 2023; Banimelhem & Amayreh, 2023, Huang, Wu, Liang, Wang & Zhao, 2023; Rinsum, 2023; Jibreel, 2023; Lee, 2023). For instance, a study by Banimelhem & Amayreh (2023) evaluated the performance of ChatGPT in translating standard Arabic sentences. The results indicated that while ChatGPT showed promise, it did not consistently outperform traditional MT systems across all tasks. However, following the rapid development of LLM, Aldawsari (2024) showed that this may not be the case with translation of Arabic colloquial expressions, highlighting the need for further research into the capabilities of LLMs in handling complex Arabic phrases.

This study aims to build on aforementioned previous research by evaluating the performance of two prominent LLMs, OpenAI’s GPT and Google’s Gemini, in translating Arabic lexical ambiguities. By comparing these LLMs with traditional MT systems such as Google Translate and SYSTRAN, the research seeks to determine whether LLMs can offer superior translation quality and accuracy for Arabic texts. The research objectives guiding this study are to assess how accurately LLMs (GPT and Gemini) translate Arabic lexical ambiguities compared to traditional MT systems (Google Translate and SYSTRAN); to identify the strengths and limitations of LLMs in handling homonyms, heteronyms, and polysemes in Arabic; and to analyze how the translation outputs of LLMs differ from those of traditional MT systems in terms of accuracy and intelligibility. By addressing these objectives, the study aims to provide valuable insights into the current capabilities of LLMs and their potential to improve the translation of Arabic texts, thereby contributing to the ongoing efforts to enhance machine translation technologies.

3. Methods

This study employs a comparative evaluation research design to assess the performance of two large language models (LLMs)—OpenAI’s

GPT and Google's Gemini—against traditional machine translation (MT) systems, specifically Google Translate and SYSTRAN. The objective is to determine the effectiveness of these LLMs in handling Arabic lexical ambiguities of homonyms, heteronyms, and polysemes.

The research instruments utilized in this study included a test suite and two questionnaires. The test suite comprised sixteen Arabic sentences (see Appendix A), which were designed to include three types of lexical ambiguity: homonyms, heteronyms, and polysemes. These different types were selected to cover a range of contexts to ensure a comprehensive evaluation of the MT systems' capabilities. Homonyms are words that are spelled and pronounced the same but have different meanings, heteronyms are words that are spelled the same but pronounced differently and have different meanings, and polysemes are words that have multiple related meanings) (Crystal, 2019). Four translation systems were evaluated in this study: OpenAI's GPT (an advanced LLM known for its extensive training on diverse text corpora), Google's Gemini (a leading LLM leveraging large-scale data and sophisticated algorithms for natural language understanding and generation), Google Translate (a widely used MT system employing neural machine translation techniques), and SYSTRAN (a traditional MT system known for its rule-based and hybrid translation methods).

The procedure began with data preparation, ensuring the test suite was consistent in format and structure. Each sentence was prepared for input into the translation systems without any modifications to the original content. Each sentence from the test suite was then translated by the four systems, and the translations were collected and recorded for further analysis (Appendices, B-E). No additional context was given to the LLM systems other than the prompt "Translate this from Arabic into English please". The translated outputs were evaluated based on accuracy (the degree to which the translated text correctly conveys the meaning of the original sentence) and intelligibility (the fluency and grammatical correctness of the translated text) (Arnold, Balkan, Meijer, Humphreys, and Sadler, 1994).

The evaluation employed both quantitative and qualitative measures. A panel of native Arabic speakers proficient in English independently assessed the translations, scoring them based on predefined criteria for accuracy and intelligibility, following the methodology outlined by Aldawsari (2023). Quantitative scoring involved averaging the intelligibility and accuracy ratings collected from the evaluators' responses to the questionnaires. Subsequently, a detailed qualitative analysis was conducted, with particular attention to sentences where the systems either excelled or faced difficulties. These cases were further examined to uncover underlying patterns and tendencies.

4. Results

This section presents the findings of the comparative evaluation of the two LLMs—OpenAI's GPT and Google's Gemini—against the traditional MT systems, Google Translate and SYSTRAN. The evaluation focused on handling Arabic homonyms, heteronyms, and polysemes (Appendices, B-E). On the whole, the evaluations demonstrate that GPT and Gemini both scored significantly higher in terms of both intelligibility and accuracy compared to Google Translate and SYSTRAN highlighting more advanced capabilities of LLMs in handling the complexities of Arabic lexical ambiguities compared to traditional MT systems.

4.1 Intelligibility of Translations

The intelligibility of the translated texts was assessed based on fluency and grammatical correctness. The results are summarized in Table 1.

Table 1. Intelligibility Scores

Sentence	GPT	Gemini	Google Translate	SYSTRAN
1	4	4	2.5	2.5
2	4	4	2.75	2.25
3	3.5	3.25	2.25	3.25
4	3.75	3.75	3	3.5
5	4	4	1.75	2.5
6	4	4	2.5	2.75
7	4	3.75	1.25	2.75
8	4	4	1.75	2.5
9	4	4	1.5	2.5
10	3.75	3.5	1.25	2.5
11	4	4	2	2.25
12	4	4	3.25	2.25
13	4	4	2.75	4
14	3.75	3.5	1.75	2.75
15	3.25	3.25	2.5	2.25
16	4	4	2.75	1.5
Average	**3.88**	**3.81**	**2.38**	**2.6**

The intelligibility scores demonstrate that GPT and Gemini consistently outperform Google Translate and SYSTRAN across all 16 sentences. GPT and Gemini achieve high scores, with average ratings of 3.88 and 3.81, respectively, indicating a high level of fluency and comprehension in translating Arabic lexical ambiguities. Google Translate's average score of 2.38 reflects moderate performance, with some translations being clear but others failing to convey the original meaning accurately. SYSTRAN, with an average score of 2.6, shows varied performance, occasionally producing intelligible translations but often struggling with the intended meaning of the source text. Overall, the data highlights the superior capability of LLMs like GPT and Gemini in handling complex linguistic features of Arabic compared to traditional MT systems.

4.2 Accuracy of Translations

After intelligibility, the accuracy of translations were assessed and results are summarized in Table 2.

Table 2. Accuracy Scores of Translations

Sentence	GPT	Gemini	Google Translate	SYSTRAN
1	4	4	2.75	3.75
2	4	4	3.75	3
3	3.75	3.5	2.5	2.5
4	4	4	2.25	1.75
5	4	4	1.25	1.25
6	4	4	1.5	1.25
7	4	3.25	1.75	1.25
8	4	4	1.5	1.75
9	4	3.25	2.5	2.75
10	4	4	3.25	2.75
11	4	3.75	3.25	2.5
12	4	4	3.75	3.25
13	4	4	1.5	4
14	3.75	3.75	2.75	2.5
15	4	4	2.5	3
16	3.75	4	3	4
Average	**3.97**	**3.84**	**2.48**	**2.58**

The accuracy scores indicate that GPT and Gemini deliver consistently higher accuracy in translating Arabic lexical ambiguities, with average scores of 3.97 and 3.84, respectively. These scores suggest that both LLMs are capable of retaining the original meaning of sentences while accurately translating them. Google Translate, with an average score of 2.48, shows variability in performance, occasionally achieving accurate translations but frequently falling short in conveying the intended message. SYSTRAN, averaging at 2.58, demonstrates a slight edge over Google Translate in some instances but still struggles significantly with maintaining accuracy in many translations. These results highlight the substantial advantage of GPT and Gemini over traditional MT systems, emphasizing their effectiveness in accurately translating complex Arabic texts.

It is clear that LLM produced significantly more accurate and readable sentences. Various examples from the translations can demonstrate this improvement. For instance, sentence 16 "سمو الأمير تركي بن متعب يزور الرياض" was translated by GPT as "Prince Turki bin Muteb visited Riyadh," which is accurate and fluent, retaining the essential elements of the original sentence with proper names correctly translated. Gemini translated it as "His Highness Prince Turki bin Mut'ib visits Riyadh," adding "His Highness," which is not explicitly in the original Arabic but can be inferred, making it more formal than GPT’s version. Google Translate's version, "Prince Turki bin tried to visit Riyadh," is incorrect as it misinterprets "يزور" (visits) as "tried to visit," altering the meaning significantly. SYSTRAN produced "His Highness of the Prince is Turkish tiring coffee forges Riyadh," which is completely incorrect and unintelligible, likely due to poor handling of named entities and idiomatic expressions. For Sentence 10 "إن أخي سعيد بحضوركم," GPT translated it as "My brother is happy with your presence," maintaining the sentiment and structure of the original sentence accurately and fluently. Gemini's version, "My brother Saeed is honored by your presence," adds a name not present in the original and changes "happy" to "honored," slightly altering the sentiment. Google Translate matched GPT’s translation with "My brother is happy with your presence," but generally demonstrated less contextual accuracy in other translations. SYSTRAN's translation, "That two brothers of happy with your attendance," misinterpreted the singular "أخي" (my brother) as plural and produced an awkward sentence. For Sentence 6 "مارسْ السباحة," GPT translated it as "Practice swimming," capturing the imperative mood of the original directly and accurately. Gemini's "He practiced swimming" changes the imperative to a declarative statement, adding "He," which is incorrect and alters the original meaning. Google Translate's "March swim" is incorrect in tense and phrasing, misinterpreting "مارسْ" (practice) as "March." SYSTRAN's "March the swimming" misinterprets the verb and produces an unnatural sentence.

An additional illustrative example of the systems' handling of polysemes is Sentence 3, "وعدٌ علي ألا أترك المثابرة," which translates as "I promise not to give up perseverance." GPT rendered this as "I promise not to give up persistence," capturing the intended meaning. Gemini’s version, "I promise myself not to give up perseverance," added an explicit reflexive pronoun, which, while grammatically correct, was not necessary and slightly changed the focus of the statement. Google Translate’s output, "Promised not to leave perseverance," was awkward and failed to convey the personal commitment implied in the original Arabic. SYSTRAN’s translation, "Promise raised not to the perseverance leaves," was incomprehensible, reflecting its inability to handle the abstract noun "المثابرة" (perseverance) in context.

In Sentence 12, "الفطر ينبت في كل مكان," (lit. the mushroom grows everywhere), GPT translated it as "Mushrooms grow everywhere," preserving both the structure and intent of the original although making mushroom plural. Gemini produced the same translation. Google Translate, however, rendered it as "The fast-breaking plants are everywhere," confusing "الفطر" (mushrooms) with its homograph meaning "breaking the fast," a term used during Ramadan. SYSTRAN’s output, "The fast-breaking plants are everywhere," repeated this error, highlighting the challenges traditional MT systems face with polysemous words, particularly when context is insufficiently accounted for.

Sentence 5, "وَعَدَ فهد أباه بالمذاكرة," provides another compelling case. Both LLM systems GPT and Gemini translated this as "Fahd promised his father to study," accurately conveying the intent of the original. In contrast, Google Translate’s version, "Fahd promised his father Palmmakrh," misinterpreted "بالمذاكرة" (to study) as an unrecognized term, resulting in a nonsensical output. SYSTRAN’s output, "Leopard of father promised him baalmdhaakrt," incorrectly interpreted "فهد" (Fahd) as "Leopard," a literal translation of the name, and failed to process the verb-object relationship, resulting in an unintelligible sentence. Finally, in Sentence 14, "زرت محافظة وادي الدواسر," meaning "I visited Wadi Al-Dawasir Province," GPT produced an accurate and fluent translation. Gemini’s output, "I visited Wadi Dawaser Governorate," was similarly precise, with "Governorate" being an acceptable alternative for "محافظة" (province). Google Translate rendered this as "I visited the Valley of Propellants province," misinterpreting "وادي الدواسر" (Wadi Al-Dawasir) and introducing an unrelated term, "propellants." SYSTRAN performed worse, producing "Governorate Valley of the Propelling visited," which failed to reflect the intended meaning and order of the original Arabic.

Overall, the analysis reveals that GPT and Gemini consistently produce higher quality translations compared to Google Translate and SYSTRAN. GPT's translations are generally more fluent and accurate, while Gemini adds a layer of formality and sometimes additional context, such as "His Highness." Google Translate struggles with idiomatic expressions and sentence structure, resulting in less intelligible translations. SYSTRAN often produces translations that are incorrect and unintelligible, likely due to poor handling of named entities and grammatical structures. The differences highlight the strengths of LLMs like GPT and Gemini in providing not only accurate but also contextually rich translations.

5. Discussion

The results indicate that GPT and Gemini offer substantial improvements over traditional MT systems in translating Arabic lexical ambiguities (Aldawsari, 2023, 2024). The higher accuracy and intelligibility scores suggest that LLMs are better equipped to handle the complexities of Arabic homonyms, heteronyms, and polysemes. However, the performance differences between GPT and Gemini were minimal, indicating that both models are similarly effective in this task.

The findings also highlight the limitations of traditional MT systems, which continue to struggle with Arabic's linguistic complexities. These results suggest a paradigm shift in how machine translation handles the complexities of language, particularly with regard to languages that have a high degree of lexical ambiguity like Arabic. Traditional MT systems have long been criticized for their inability to accurately translate semantically complex expressions. The significant improvement demonstrated by GPT and Gemini showcases the potential of LLMs to bridge this gap, providing translations that accurately render the intended meaning.

The disparity between the limitations identified in previous studies (e.g., Al-Kharabsheh & Yassin, 2017; At-tall, 2019) and the potential for LLMs suggested by this study, alongside more recent research (e.g., Aldawsari, 2024; Boughorbel & Hawasly, 2023; Banimelhem & Amayreh, 2023), hints at progress in machine translation, particularly within LLM-driven approaches. While recent research (e.g., Banimelhem & Amayreh, 2023) found that LLM systems performed below average compared to established MT systems on standard Arabic translation, this study suggests that LLMs may produce more accurate renditions of lexically ambiguous Arabic sentences. This is supported by Aldawsari’s (2024, p.237) findings where there was an “advantage” of LLMs in dealing with “non-standard” Arabic expressions. The findings indicate that LLMs have the potential to overcome the challenges associated with Arabic lexical ambiguities, offering accurate translations.

Despite the advancements showcased by both GPT and Gemini, their remarkably close performance suggests that the field's progress is primarily driven by broader improvements in training techniques rather than model-specific breakthroughs. As such, further improvements in this area may benefit from a focus on refining these common methodologies and exploring fine-tuning methods that can enhance translation output. Moreover, the continued struggles of traditional MT systems call attention to a fundamental limitation in their approach to language processing. These systems, often based on statistical and rule-based methods, lack the sophisticated contextual understanding that LLMs bring. The quality achieved by GPT and Gemini highlight their ability to generate translations that are more fluent and closer to human-level understanding. This improvement is particularly evident in handling idiomatic expressions and culturally specific phrases (Aldawsari, 2024), which are critical for accurate and meaningful translations.

The implications of these findings are far-reaching. The findings indicate that LLMs have the potential to overcome the challenges associated with challenging Arabic phrases, offering accurate translations. This has several important implications for the field of machine translation and beyond. For instance, the demonstrated capabilities of LLMs suggest a shift in the development priorities for translation technologies. Developers and researchers should focus on integrating LLM-based approaches into existing translation workflows, especially for linguistically complex languages. By leveraging the contextual and semantic understanding of LLMs, these systems could significantly improve translation quality. In addition, these findings emphasize the importance of tailoring translation technologies to handle the specific linguistic features of target languages. For instance, Arabic’s rich morphology and frequent lexical ambiguities present unique challenges that traditional systems struggle to address. By contrast, LLMs hold significant promise in this area, demonstrating the potential for machine translation systems to be fine-tuned for specific linguistic contexts. The improvements observed in LLM performance highlight opportunities for hybrid translation models that combine the scalability of traditional MT systems with the accuracy and fluency of LLMs. For example, traditional MT systems could handle simpler translation tasks while delegating contextually complex or ambiguous sentences to LLMs, resulting in more robust and efficient systems. Finally, as LLMs become more sophisticated, the skillsets required of translators and linguists will also evolve, highlighting the need for curricula that integrate training in AI-assisted tools, enabling professionals to collaborate effectively with advanced technologies.

In conclusion, while GPT and Gemini represent a significant step forward in the translation of Arabic lexical ambiguities, there remains a need for ongoing research and development to fully realize the potential of LLMs. Traditional MT systems still play a role in the broader landscape of machine translation, but their limitations must be addressed through innovative approaches that leverage the strengths of LLMs. The future of machine translation lies in the continued integration of advanced AI technologies, pushing the boundaries of what is possible in language processing and translation.

6. Conclusion

This study set out to evaluate the performance of large language models (LLMs), specifically OpenAI’s GPT and Google’s Gemini, in translating Arabic lexical ambiguities. By comparing these models with traditional machine translation (MT) systems such as Google Translate and SYSTRAN, the study aimed to determine whether LLMs can offer superior translation accuracy and intelligibility. The results demonstrate that GPT and Gemini significantly outperform traditional MT systems in handling Arabic homonyms, heteronyms, and polysemes. The higher accuracy and intelligibility scores achieved by the LLMs indicate their advanced capability in managing the complexities of Arabic lexical ambiguities. Despite the clear advantages of LLMs, the performance differences between GPT and Gemini were minimal, suggesting that both models are equally effective in this context. This finding points to the high quality of current LLM architectures and training methodologies, highlighting the significant progress made in natural language processing. The findings have important implications for the development and deployment of machine translation systems. For translators, integrating LLMs into MT workflows can lead to more accurate and easily readable translations, particularly for complex languages such as Arabic. For developers, the results emphasize the value of investing in advanced LLM technologies to enhance translation quality.

6.1 Limitations and Further Research

While this study offers valuable insights into the capabilities of LLMs in handling Arabic lexical ambiguities, it is important to acknowledge several limitations that may effect the scope of the findings. First, the test suite used in this study consisted of only sixteen sentences, which, while carefully selected to represent homonyms, heteronyms, and polysemes, may not fully capture the diversity and complexity of Arabic lexical ambiguity across various contexts and dialects. A larger and more varied dataset would provide a more comprehensive evaluation of the systems’ performance and highlight additional strengths and weaknesses. Future studies should expand the test suite to include larger corpora, covering a broader range of linguistic phenomena and regional dialects to better assess the robustness of LLMs across different contexts. Second, the evaluation relied primarily on human judgment through questionnaires to assess intelligibility and accuracy. While this approach ensures qualitative insights, it may introduce bias. Future research could complement human evaluation with automated metrics to provide a more objective and reproducible analysis. Additionally, a larger panel of evaluators with diverse linguistic backgrounds could help mitigate bias and enhance the reliability of the findings. Finally, this study focused exclusively on one language, Arabic, limiting the applicability of the findings to other languages with similar linguistic features. While Arabic’s rich morphology and syntactic complexity provide a rigorous testbed, further research is needed to explore how LLMs perform on other languages. Cross-linguistic studies would help determine whether the observed advantages of LLMs are universal or language-specific. Future research could also evaluate the systems' performance on longer texts to assess how well LLM systems utilize contextual information for accurate translations.

In light of these limitations, several promising avenues for future research emerge. Investigating hybrid approaches that combine the strengths of LLMs with traditional MT systems could yield practical solutions for improving translation accuracy and practicality. Further studies could also track the evolution of LLM performance over time, particularly as newer models and training techniques are developed. Finally, exploring fine-tuning methods for domain-specific applications—such as legal translations—could help maximize the utility of LLMs in specialized contexts. In summary, the adoption of LLMs like GPT and Gemini represents a significant advancement in the field of machine translation, particularly for complex languages such as Arabic. While traditional MT systems continue to play a role, their limitations highlight the need for ongoing innovation and research. By exploring new methodologies and expanding the scope of evaluation, future research can further enhance the capabilities of machine translation systems, ultimately contributing to more effective and accurate language processing.

Acknowledgments

I sincerely thank Prince Satam bin Abdulaziz University for funding this study. I extend my gratitude to the evaluators for their invaluable contributions.

Authors’ contributions

Not applicable.

Funding

This study is supported via funding from Prince Sattam bin Abdulaziz University project number (PSAU/2024/R/1446).

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal’s policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

Open access

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

References

Aldawsari, H. A. H. (2023). Comparing the Performance of Google Translate and SYSTRAN on Arabic Lexical Ambiguity. *Arab World English Journal for Translation & Literary Studies*, *7*(3), 19-34. https://doi.org/10.24093/awejtls/vol7no3.2

Aldawsari, H. A. H. (2024). Evaluating Translation Tools: Google Translate, Bing Translator, and Bing AI on Arabic Colloquialisms. *Arab World English Journal*, *Special Issue*, pp. 237–251. https://doi.org/10.24093/awej/chatgpt.16

Ali, M. A. (2020). Quality and machine translation: An evaluation of online machine translation of English into Arabic texts. *Open Journal of Modern Linguistics*, *10*(05), 524-548. https://doi.org/10.4236/ojml.2020.105030

Al-Jarf, R. (2023). Pluralization of social media loan terminology in colloquial Arabic. *British Journal of Applied Linguistics*, *3*(2), 60-68. https://doi.org/10.32996/bjal.2023.3.2.6

Al-Kharabsheh, A., & Yassin, O. (2017). Translation of colloquialisms in the Arabic-into-English subtitled film, the dupes. *International Journal of Comparative Literature and Translation Studies*, *5*(3), 18. https://doi.org/10.7575/aiac.ijclts.v.5n.3p.18

Almahasees, Z. (2021). *Analysing English-Arabic Machine Translation: Google Translate, Microsoft Translator and Sakhr*. Routledge. https://doi.org/10.4324/9781003191018

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

Arnold, D., Balkan, L., Meijer, S., Humphreys, R. L., & Sadler, L. (1994). *Machine Translation: An Introductory Guide*. Manchester: NCC Blackwell.

At-tall, S. M. (2019). *Comparative study between Google Translator and human translator in rendering colloquial Arabic expressions in the late Prime Minister Wasfi At-Tall's speeches into English*, (Unpublished Master's thesis). Yarmouk University, Irbid, Jordan.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). *Neural Machine Translation by Jointly Learning to Align and Translate*. International Conference on Learning Representations. Retrieved from https://arxiv.org/pdf/1409.0473

Banimelhem, O., & Amayreh, W. (2023). Is ChatGPT a Good English to Arabic Machine Translation Tool?. In *2023 14th International Conference on Information and Communication Systems* (ICICS) (pp. 1-6). IEEE. https://doi.org/10.1109/ICICS60529.2023.10330525

Boughorbel, S., & Hawasly, M. (2023). Analyzing Multilingual Competency of LLMs in Multi-Turn Instruction Following: A Case Study of Arabic. *SIGARAB Arabic NLP*. https://doi.org/10.18653/v1/2023.arabicnlp-1.11

Crystal, D. (2019). *The Cambridge encyclopedia of the English language* (3rd ed.). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/9781108528931

Habash, N., & Diab, M. T. (2012). Arabic natural language processing: Why it matters and how to get started. In *Proceedings of the ACL-HLT 2012 Workshop on Computational Approaches to Arabic Script Languages* (pp. 1-9). Association for Computational Linguistics.

Harrat, S., Meftouh, K., & Smaili, K. (2019). Machine translation for Arabic dialects (survey). *Information Processing & Management*, *56*(2), 262-273. https://doi.org/10.1016/j.ipm.2017.08.003

Huang, H., Wu, S., Liang, X., Wang, B., & Zhao, T. (2023). Towards Making the Most of LLM for Translation Quality Estimation. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 375-386). Cham: Springer Nature Switzerland. https://doi.org/10.3115/1073083.1073135

Hutchins, W. J. (2001). *Early Years in Machine Translation*. John Benjamins Publishing Company. https://doi.org/10.1075/sihols.97

Hutchins, W., & Somers, H. (1992). *An introduction to machine translation*. London: Academic Press.

Jibreel, I. (2023). Online machine translation efficiency in translating fixed expressions between English and Arabic (proverbs as a case-in-point). *Theory and Practice in Language Studies*, *13*(5), 1148-1158. https://doi.org/10.17507/tpls.1305.07

Koehn, P. (2010). *Statistical machine translation*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Lee, T. (2023). Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review*, *14*(4), 60-66. https://doi.org/10.1515/applirev-2023-0122

Okpor, M. D. (2014). Machine translation approaches: Issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, *11*(5), 159. https://doi.org/10.20943/01201702.5457

Rinsum, H. (2023). CHATGPT, Bing Chat or bard? *Agrarzeitung, 78*(39), 10-10. https://doi.org/10.51202/1869-9707-2023-39-010

Turovsky, B. (2016). *Found in translation: More accurate, fluent sentences in Google Translate*. Retrieved from https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic Machine Translation: A Survey with Challenges and Future Directions. *IEEE Access*, *9*, 161445-161468. https://doi.org/10.1109/ACCESS.2021.3132488

Appendices

Appendix A

Arabic Test Suite

Table 1. Arabic Sentences for the test suite

1. وصلت إلى الفرع في الساعة الخامسة.
2. الفرع ثقيل جداً.
3. وعدٌ علي ألا أترك المثابرة.
4. وعدَّ المبلغ كاملاً أمام أصدقائه.
5. وَعَدَ فهد أباه بالمذاكرة.
6. مارسَ خالد السباحة.
7. مارسْ السباحة.
8. ذهبَ محمد إلى جدته.
9. ذَهَبُ محمد غالي الثمن.
10.إن أخي سعيد بحضوركم .
11.أخي سعيد ذهب إلى المدرسة.
12.الفطر ينبت في كل مكان.
13.عيد فطر سعيد.
14.زرت محافظة وادي الدواسر.
15.التعامل مع هذا الرجل متعب.
16.سمو الأمير تركي بن متعب يزور الرياض.

Appendix B

GPT Output

Table 2. English translations by Open AI’s GPT

1. I arrived at the branch at five o'clock.
2. The branch is very heavy.
3. I promise not to give up persistence.
4. He counted the full amount in front of his friends.
5. Fahd promised his father to study.
6. Khaled practiced swimming.
7. Practice swimming.
8. Muhammad went to his grandmother.
9. Muhammad's gold is expensive.
10. My brother is happy with your presence.
11. My brother Saeed went to school.
12. Mushrooms grow everywhere.
13. Happy Eid al-Fitr.
14. I visited Wadi Al-Dawasir Province.
15. Dealing with this man is tiring.
16. Prince Turki bin Muteb visited Riyadh.

Appendix C

Gemini Output

Table 3. English translations by Google’s Gemini

1. I arrived at the branch at five o'clock.
2. The branch is very heavy.
3. I promise myself not to give up perseverance.
4. He counted the full amount in front of his friends.
5. Fahd promised his father to study.
6. Khalid practiced swimming.
7. He practiced swimming. (Note: The subject "He" is implied in Arabic but needs to be added in English for clarity.)
8. Muhammad went to his grandmother.
9. Muhammad's gift is very expensive.
10. My brother Saeed is honored by your presence.
11. My brother Saeed went to school.

12. Mushrooms grow everywhere.
13. Happy Eid al-Fitr.
14. I visited Wadi Dawaser Governorate.
15. Dealing with this man is tiring.
16. His Highness Prince Turki bin Mut'ib visits Riyadh.

Appendix D

Google Translate Output

Table 3. English translations by Google Translate

1. Arrived at the branch in the fifth hour.
2. Branch is too heavy.
3. Promised not to leave perseverance.
4. Promised the full amount in front of his friends.
5. Fahd promised his father Palmmakrh.
6. March Khalid swimming.
7. March swim.
8. Muhammad went to his grandmother.
9. Mohammed went expensive.
10. My brother is happy with your presence.
11. Said brother went to school.
12. Mushrooms sprouting everywhere.
13. Mushroom happy holiday.
14. I visited the Valley of Propellants province.
15. Dealing with this tired man.
16. Prince Turki bin tried to visit Riyadh.

Appendix E

SYSTRAN Output

Table 4. English translations by SYSTRAN

1. His Highness of the Prince is Turkish tiring coffee forges Riyadh.
2. The dealing with this man is tiring.
3. Governorate Valley of the Propelling visited.
4. Happy Eid Al Fitr.
5. The fast-breaking plants are everywhere.
6. Two brothers are happy gold to the teacher.
7. That two brothers of happy with your attendance.
8. Muhammad expensive went.
9. Muhammad went to be new to him.
10. March the swimming.
11. Immortal swimming March.
12. Leopard of father promised him baalmdhaakrt.
13. Promise informed as hope in front of his friends.
14. Promise raised not to the perseverance leaves.
15. The branch is heavy very.
16. The branch in the hour arrived to fifth.