OSACT 2024 Task 2: Arabic Dialect to MSA Translation

Hanin Atwany[1], **Nour Rabih**[1], **Ibrahim Mohammed**[1], **Abdul Waheed**[1], **Bhiksha Raj**[1,2]
[1]Mohamed bin Zayed University of Artificial Intelligence,
[2]Carnegie Mellon University
{hanin.atwany, nour.rabih, ibrahim.mohammed,abdul.waheed,bhiksha.raj}@mbzuai.ac.ae

Abstract

We present the results of Shared Task "Dialect to MSA Translation", which tackles challenges posed by the diverse Arabic dialects in machine translation. Covering Gulf, Egyptian, Levantine, Iraqi and Maghrebi dialects, the task offers 1001 sentences in both MSA and dialects for fine-tuning, alongside 1888 blind test sentences. Leveraging GPT3.5, a state-of-the-art language model, our method achieved a BLEU score of 29.61. This endeavor holds significant implications for Neural Machine Translation (NMT) systems targeting low-resource languages with linguistic variation. Additionally, negative experiments involving fine-tuning AraT5 and No Language Left Behind (NLLB) using the MADAR Dataset resulted in BLEU scores of 10.41 and 11.96, respectively. Future directions include expanding the dataset to incorporate more Arabic dialects and exploring alternative NMT architectures to further enhance translation capabilities.

1. Introduction

Arabic, a language spoken by over 420 million people globally, boasts a rich tapestry of dialectal variations. This linguistic landscape comprises both Modern Standard Arabic (MSA), the formal variant employed in official domains such as government communications, national media, and education, and a myriad of regional dialects used predominantly in everyday interactions (Harrat et al., 2017). The differences between these dialects, which range from subtly distinct to completely unintelligible (Abdul-Mageed et al., 2022), pose a formidable challenge for machine translation systems.

Historically, the focus of machine translation systems has been predominantly on MSA. This causes those systems to struggle to capture the intricate differences inherent in dialects. Consequently, achieving accurate translation between these linguistic variants remains paramount. Addressing this challenge is crucial to enhance communication and comprehension within the Arabic-speaking world.

In the field of Natural Language Processing (NLP), dialect identification and translation are two critical areas of research. This paper concentrates on the latter, specifically examining the performance of various models in translating sentences from diverse Arabic dialects into MSA. This investigation is set in the context of the second shared task at The 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT6), which aims to address the complexities of dialect translation.

In particular, we examined the performance of three distinct methods. Firstly, we fine-tuned the AraT5 transformer model (Nagoudi et al., 2022) using diverse corpora sourced from MADAR (Bouamor et al., 2018a). Secondly, we explored the inference capabilities of the NLLB model (Costa-jussà et al., 2022). Lastly, we employed a prompting technique with GPT3.5 to facilitate dialect-to-MSA translation. By comparing these three methods, we aim to evaluate the strengths and limitations of each approach and identify the most effective solution for dialect-to-MSA translation. Our investigation provides valuable insights into the challenges of dialect translation and highlights the potential of state-of-the-art language models in addressing these challenges.

Task 2: Dialect to MSA Machine Translation
The objective of this task is to develop a model that converts Arabic from five (Gulf, Egyptian, Levantine, Iraqi, and Maghrebi) dialects to MSA. Participants can use any resources available to develop their systems.

2. Related Work

Over the past decade, advancements in the field of dialect to Modern Standard Arabic translation have been notable, driven by the imperative to foster communication and comprehension across varied Arabic dialects and the standardized form of the language (Mohamed et al., 2024). Despite these strides, challenges persist in achieving high-quality translations (Abdelali et al., 2024).

A study by (Al-Sabbagh, 2024) scrutinized the performance of Google Translate in translating Egyptian Arabic adjuncts, revealing low BLEU scores and various issues, including literal translations of idiomatic adjuncts and misinterpretation of dialectal adjuncts.

In addressing the translation challenges within NMT systems for Arabic dialects, (Moukafih et al.,

OSACT 2024 Workshop, pages 98–103

2021) investigated multitasking learning strategies, yielding noteworthy enhancements in BLEU scores for Algerian Modern Standard Arabic and Moroccan Palestinian dialects.

Recent efforts have focused on developing models that deal with translating single dialects to MSA. For instance, (Sghaier and Zrigui, 2020) proposed a rule-based machine translation system for translating Tunisian dialect to MSA, achieving a BLEU score of 55.22. Furthermore, (Sallam and Mousa, 2024) assessed the performance of AI chatbot ChatGPT in responding to health queries in Tunisian and Jordanian Arabic dialects. Their study revealed that GPT-4 exhibits slightly better performance than ChatGPT[1], with above-average scores in Jordanian Arabic but average scores in Tunisian Arabic. However, responses in both dialects fell significantly short compared to English, emphasizing the importance of linguistic and cultural diversity in AI model development, particularly in healthcare.

A comprehensive evaluation conducted by (Kadaoui et al., 2023) assessed Bard and ChatGPT for machine translation across ten Arabic varieties, encompassing Classical Arabic (CA), MSA, and country-level dialectal variants. Their findings indicated that Large Language Models (LLMs) may face challenges with dialects possessing minimal public datasets but generally outperform existing commercial systems in dialect translation. However, instruction-tuned LLMs still trail behind commercial systems like Google Translate in CA and MSA translation. Their human-centric study also underscored Bard's limited ability to adhere to human instructions in translation contexts.

In conclusion, these studies underscore the necessity for continued research and development aimed at enhancing the linguistic inclusivity of LLMs and addressing the distinctive hurdles associated with translating diverse dialects to MSA.

3. Data

3.1. Shared Task Data

We conducted thorough evaluations on both the validation and test sets provided for this shared task.

3.1.1. Validation Dataset

The validation dataset provided in this shared task, comprised a total of 1001 source-to-target examples, evenly distributed among dialects as follows: 200 Egyptian, 200 Maghrebi, 200 Levantine, 201 Gulf, and 200 Iraqi examples.

Notably, some examples featured Arabised text, where English words were transcribed using Arabic letters, as demonstrated below:

{ "id": 419221,
"dialect": "Iraqi",
"source": "يس آي لاف يو تو ماتش تعالي وين نص دينار",
"target": "نعم أنا أحبك كثيرًا تعالي أين نصف دينار"
"English translation": "Yes, I love you too much, come where is half a Dinar"

In this instance, the source sentence incorporates English phrases represented in Arabic script, while the corresponding target sentence reflects the translation into Modern Standard Arabic. Such instances posed unique challenges during evaluation and were included in the validation dataset to assess translation quality comprehensively.

Moreover, the validation dataset includes 22 sentences with a length greater than 128 characters, further enriching the evaluation process and highlighting the model's ability to handle complex linguistic structures

3.1.2. Test Datasets

The test dataset, comprised 1888 examples, each presenting its own unique linguistic challenge. These examples were distributed across different dialects as follows: 314 Egyptian, 343 Maghrebi, 568 Levantine, 77 Iraqi, and 586 Gulf.

The source sentences provided cover a broad spectrum of topics and linguistic structures, reflecting the rich diversity of Arabic dialects. They encompass both everyday conversational phrases and more formal expressions, offering a comprehensive representation of language usage in real-world scenarios.

Among these sentences, 45 exceed a length of 128 characters, presenting additional complexity to the translation task. Furthermore, the dataset includes instances of words with repeated characters, as exemplified by:

{ "مسااااء النور والسرووور وحنا نوحشنااااك بزاااااف عمري" }

Despite these challenges, the diversity in content and language enriches the dataset, enabling a thorough evaluation of the model's proficiency in handling various linguistic features and contexts.

3.2. Finetuning Dataset

The MADAR Arabic Dialect Corpus and Lexicon (Bouamor et al., 2018a), utilized in our study to fine-tune the models, represents a comprehensive resource designed to facilitate research in machine translation, particularly focusing on the translation challenges presented by Arabic dialects. The dataset consists of 25 parallel translations for 25 cities having 2,000 sentences each, in addition to their MSA equivalents and is divided into training, development, and test sets. This dataset is

[1]We refer GPT-3.5 as ChatGPT in our work.

Dialect Region	Cities Included
Egyptian	Cairo, Alexandria
Gulf	Doha, Jeddah, Muscat, Riyadh
Iraqi	Baghdad, Basra, Mosul
Levantine	Aleppo, Amman, Beirut, Damascus, Jerusalem, Salt
Maghreb	Algiers, Fes, Rabat, Sfax, Tunis

Table 1: MADAR Dataset Dialect Divisions

instrumental in understanding the linguistic diversity across the Arabic-speaking world, featuring a collection of text samples from a wide array of cities, each with its unique dialectical characteristics. For the purpose of our experiments, the dataset was meticulously organized into five distinct groups (as specified by the task), each representing a major geographical and dialectal region within the Arab world. This division was used in dialect specific finetuning.

4. Methodology

4.1. Supervised Models

NLLB. NLLB model is designed to bridge language gaps by extending translation support to a wide array of languages, with a particular focus on those with limited resources. It employs an innovative conditional compute model based on the Sparsely Gated Mixture of Experts framework, along with curated datasets and training techniques tailored for low-resource languages. In our assessment, we evaluated the NLLB 3.3B model in two scenarios: with fine-tuning on the development dataset and without fine-tuning on the test dataset.

Supervised NLLB. We finetuned NLLB 3.3B. Utilizing the MADAR Parallel Corpus Dataset, which contains data from various Arabic dialects translated into Modern Standard Arabic (MSA) (Bouamor et al., 2018a).

AraT5. AraT5 is a state-of-the-art language model specifically designed for understanding and generating Arabic text. Building upon the T5 (Text-to-Text Transfer Transformer) architecture (Raffel et al., 2023), which treats every text-based task as a "text-to-text" problem, AraT5 is fine-tuned to excel in processing and generating Arabic content across a wide range of tasks. These include text summarization, question answering, text classification, and translation. The model has been trained on a diverse corpus of Arabic text, enabling it to grasp the nuances of the language, including its dialects and classical forms.

Supervised AraT5. In the initial phase of our experiments, the model was deployed for translation tasks without any prior fine-tuning. This approach, however, did not yield successful outcomes in generating translations, primarily attributable to the constraints of the model's training. Specifically, the model was architected to facilitate machine translation from dialectal Arabic to English, with no inherent training to support translation from various dialects into Modern Standard Arabic (Nagoudi et al., 2022).

To address this, a subsequent stage of fine-tuning was implemented, utilizing the MADAR dataset as a foundational corpus. This dataset was anticipated to enhance the model's dialectal comprehension and translation efficacy. However, the results fell short of expectations, which revealed a lower than anticipated BLEU score.

AraT5-finetuned dialect-specific. Recognizing the need for a more tailored approach to capture the characteristics of each Arabic dialect, the models were fine-tuned separately for each specific dialect contained in this task. This refined strategy was predicated on the hypothesis that dialect-specific fine-tuning would enable the model to more accurately learn and replicate the unique linguistic features and idiomatic expressions inherent to each dialect. This method was designed to fix the early problems the model had when trying to translate in a general way. By doing this, we hoped to make the translations better overall and get higher scores on translation quality tests (BLEU scores).

4.2. Zero-Shot Models

We evaluate GPT3.5 (Ouyang et al., 2022) extensively to translate various dialects into modern standard Arabic. Especially, we evaluate GPT3.5 in zero and few-shot settings. We choose three examples in the few-shot setting as (Kadaoui et al., 2023) show it as the optimal setting across a wide range of Arabic to English translation tasks. We provide more details about our prompt in Table 2.

Zero-shot. We evaluate GPT3.5 in a zero-shot setting with a simple prompt asking the model to translate dialectal Arabic into MSA. We provide the zero-shot prompt template in Table 2.

Few-Shot. We also use GPT3.5 in the 3-shot setting by providing three examples from each dialect. We keep the example static throughout the dialect. Our 3-shot prompt can be found in Table 2.

Few-Shot with Self-Correction. We find that despite providing examples there seem to be issues with the translation. To address this issue, we experiment with a modified prompt that asks the model to find its mistakes and correct itself. We provide a step-by-step guide to do the task. Our refinement process improves our score by approximately

Shot	Prompt
Zero-shot	Translate the given input text from {dialect} Arabic dialect into Modern Standard Arabic (MSA). {dialect}:{input} MSA: []
Few-Shot	Translate the following input text from {dialect} Arabic dialect into the Modern Standard Arabic (MSA). The output should be in Arabic script only. Here are some examples: {examples} {dialect}:{input} MSA: []
Few-Shot with Self-Correction	Following is the Modern Standard Arabic (MSA) translation from {dialect} Arabic. {dialect}: {input} MSA: {msa} Please correct the MSA translation for the input in {dialect}. An accurate translation should consist solely of Modern Standard Arabic (MSA) words and accurately translate the given input. Here are some examples: {examples} Here is a step-by-step guide to do the task: 1. Identify any mistakes in the translation. 2. Correct the mistakes by replacing them with the correct MSA words or phrases. 3. Provide the final corrected MSA translation. Generate only the corrected MSA translation; no additional information is needed. If no changes are required, then produce the same translation. {dialect}: {input} Corrected MSA: []

Table 2: Zero-shot, few-shot, and self-correcting prompt templates. We format the prompt with appropriate input and examples before feeding it to ChatGPT.

2 points in terms of BLEU score. We report our self-correction prompt in Table 2.

4.3. Experimental Setup

We initially explored the efficacy of zero-shot prompting for Arabic dialect-to-Modern Standard Arabic translation tasks. While zero-shot prompting of GPT3.5 provided a solid baseline, we further investigated the impact of increasing the prompt complexity through a three-shot prompting approach. Remarkably, our experiments revealed a substantial improvement in BLEU scores when transitioning from zero-shot to three-shot prompting. By incorporating additional context and refining the prompts, the model gained a deeper understanding of the translation task, resulting in more accurate and fluent translations.

5. Results

BLEU score obtained using several models is recorded in Table 3. Our results show that GPT3.5 outperformed the other models in dialectal Arabic to MSA translation, with a BLEU score of 29.61.

The NLLB 3.3B Base model achieved a BLEU score of 11.96. However, the fine-tuned NLLB yielded a BLEU score lower than that of the base NLLB model without fine-tuning of 9.00.

There could be several reasons for this unexpected result:

- Heterogeneous dataset: Fine-tuning the NLLB model on the entire dataset while specifying the source language as "arb_Arab" is inaccurate, considering the dialectal variations within the MADARA dataset.

  The MADAR dataset is diverse, comprising data from multiple Arabic dialects, which may have contributed to a decline in performance

owing to the substantial differences among each dialect.

- Lack of dialect-specific fine-tuning: The fine-tuning process did not involve separate fine-tuning for each dialect. This could have led to the model being unable to learn the specific characteristics of each dialect, resulting in a lower BLEU score.

On the other hand, the AraT5 fine-tuned model achieved a BLEU score of 9.41 across all dialects. However, when fine-tuned specifically for each dialect, there was a notable improvement, with a BLEU score of 10.41.

These results suggest that GPT3.5 is more effective in capturing the features of dialectal Arabic and translating them into MSA compared to the other models.

The lower BLEU scores for the NLLB 3.3B Base and AraT5 finetuned models may be due to the complexity and variability of dialectal Arabic, which can make it challenging to generalize from the training data.

The highest BLEU scores were achieved through iterative improvements to the prompting strategy applied to GPT-3.5. Initially, the model's performance was enhanced by incorporating examples of dialect-to-Modern Standard Arabic translations into the prompt, resulting in a BLEU score of 28. Subsequently, further refinement was achieved by integrating step-by-step instructions for self-correction within the prompt framework. This iterative approach culminated in the attainment of the highest BLEU score on the test dataset, reaching 29.61.

Model	BLEU
NLLB-3.3B finetuned	9.00
AraT5 finetuned	9.41
AraT5-finetuned dialect-specific	10.41
NLLB-3.3B	11.96
ChatGPT (0-shot)	21.84
ChatGPT (3-shot)	28.00
ChatGPT (3-shot) with self-Correction	29.61

Table 3: BLEU score on the *Test* dataset.

6. Conclusion

Our experiments highlight the challenges in dialectal Arabic to MSA translation, particularly in dealing with heterogeneous datasets and the importance of dialect-specific fine-tuning. Our results also demonstrate the potential of using state-of-the-art language models like GPT to improve translation performance. Future work could involve exploring different fine-tuning strategies such as the mixture of experts to improve the BLEU score further.

7. Bibliographical References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, et al. 2024. Larabench: Benchmarking arabic ai with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rania Al-Sabbagh. 2024. The negative transfer effect on the neural machine translation of egyptian arabic adjuncts into english: The case of google translate. *International Journal of Arabic-English Studies*, 24(1):95–118.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018a. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018b. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2017. Machine translation for arabic dialects (survey). *Information Processing Management*, 56:262–273.

Karima Kadaoui, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. *arXiv preprint arXiv:2308.03051*.

Yasir Abdelgadir Mohamed, Akbar Khanan, Mohamed Bashir, Abdul Hakim HM Mohamed, Mousab AE Adiel, and Muawia A Elsadig. 2024. The impact of artificial intelligence on language translation: A review. *IEEE Access*, 12:25553–25579.

Youness Moukafih, Nada Sbihi, Mounir Ghogho, and Kamel Smaïli. 2021. Improving machine translation of arabic dialects through multi-task learning. In *International Conference of the Italian Association for Artificial Intelligence*, pages 580–590. Springer.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. Arat5: Text-to-text transformers for arabic language generation.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Malik Sallam and Dhia Mousa. 2024. Evaluating chatgpt performance in arabic dialects: A comparative study showing defects in responding to jordanian and tunisian general health prompts. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, 2024:1–7.

Mohamed Ali Sghaier and Mounir Zrigui. 2020. Rule-based machine translation from tunisian dialect to modern standard arabic. *Procedia Computer Science*, 176:310–319. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 24th International Conference KES2020.