Arabic Train at NADI 2024 shared task: LLMs' Ability to Translate Arabic Dialects into Modern Standard Arabic

Anastasiia Demidova, Hanin Atwany, Nour Rabih, Sanad Shaban
Mohamed bin Zayed University of Artificial Intelligence,
{anastasiia.demidova, hanin.atwany, nour.rabih, sanad.shaban}@mbzuai.ac.ae

Abstract

Navigating the intricacies of machine translation (MT) involves tackling the nuanced disparities between Arabic dialects and Modern Standard Arabic (MSA), presenting a formidable obstacle. In this study, we delve into Subtask 3 of the NADI shared task (Abdul-Mageed et al., 2024), focusing on the translation of sentences from four distinct Arabic dialects into MSA. Our investigation explores the efficacy of various models, including Jais, NLLB, GPT-3.5, and GPT-4, in this dialect-to-MSA translation endeavor. Our findings reveal that Jais surpasses all other models, boasting an average BLEU score of 19.48 in the combination of zero- and few-shot setting, whereas NLLB exhibits the least favorable performance, garnering a BLEU score of 8.77.

1 Introduction

The diverse array of Arabic dialects poses a significant challenge for machine translation (MT) systems aiming to connect the different spoken varieties with Modern Standard Arabic, the standardized written form of the language. In recent years, there has been a notable surge of interest in developing effective MT systems capable of accurately translating Arabic dialects into MSA. Such advancements not only foster smoother communication across diverse Arabic-speaking communities but also expand access to a wide array of Arabic content, enriching cultural exchange and understanding.

In our endeavor to confront this challenge and foster progress in Arabic dialect MT, we embarked on an exploration within the framework of the NADI shared task. This task specifically focuses on the sentence-level translation of four major Arabic dialects - Egyptian, Emirati, Jordanian, and Palestinian - into MSA.

Our primary objective lies in evaluating and comparing the performance of three distinct approaches, GPT, Jais, and NLLB, in the task of translating Arabic dialects into MSA at the sentence level. In addition, Google Translate assesses the performance of translating dialectal text into English and French, and subsequently translating it back into Arabic (MSA). To facilitate this evaluation, we meticulously curated our own datasets for model refinement. Our approach involved Google Translate translations to generate parallel datasets along with the gold MSA translations. This dataset shifts the translation task to a translation-correction task.

Subsequently, all models underwent a rigorous evaluation using the test data set, which includes samples drawn from the four specified dialects. Our observations yielded noteworthy insights, with the performance analysis revealing that Jais exhibited superior translation proficiency compared to GPT, NLLB, and Google Translate as evidenced by higher BLEU scores.

These findings underscore the importance of continued exploration and innovation in the realm of Arabic dialect MT, as we strive to develop robust systems that not only bridge linguistic divides, but also foster cross-cultural exchange and understanding. Through collaborative efforts and thoughtful refinement, we aim to chart a path towards more inclusive and accessible language technologies for Arabic speakers worldwide

2 Related Work

The landscape of Arabic dialect Machine Translation (MT) has undergone significant evolution, transitioning from earlier rule-based and statistical methods to more sophisticated neural network models. Sawaf (2010) laid the groundwork by exploring these traditional approaches together. However, with the advent of deep learning, neural network-based methods have taken center stage.

Recent studies have delved into the capabilities

Proceedings of The Second Arabic Natural Language Processing Conference, pages 729–734
August 16, 2024 ©2024 Association for Computational Linguistics

of Large Language Models (LLMs) in translating Arabic dialects. Kadaoui et al. (2023) conducted an extensive investigation into the translation proficiency of instruction-tuned LLMs across various Arabic varieties. Their research illuminated the challenges faced by LLMs like Bard and ChatGPT when dealing with dialects lacking substantial public datasets. Nevertheless, these models demonstrated remarkable performance, surpassing existing commercial systems and affirming their effectiveness as dialect translators.

Conversely, Farhan et al. (2020) pioneered unsupervised dialectal Neural Machine Translation (NMT), presenting two systems for this novel challenge. The Dialectal to Standard Language Translation (D2SLT) system leverages dialectal similarities via unique strategies, achieving promising results. Notably, the highest BLEU score in the unsupervised setting reached 32.14, at bar with the best score obtained in the supervised setting, which stands at 48.25.

Many proposed methods depend on accurately identifying the dialect before translation, posing several challenges in the process. Keleg and Magdy (2023) address the limitations of current Automatic Arabic Dialect Identification (ADI) systems in distinguishing between Arabic micro-dialects. They critique the prevailing approach of treating ADI as a single-label classification problem, arguing that it contributes to system failures. Through manual error analysis conducted by seven native Arabic speakers, approximately 66% of identified errors were found to be inaccuracies. Consequently, the paper proposes reframing ADI as a multi-label classification task and offers recommendations for enhancing future ADI datasets.

This paper focuses on exploring the capabilities of the most recent models, such as Jais, GPT-3.5, and GPT-4, for dialect to MSA translation.

3 Data

3.1 Shared Task Data

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

3.1.1 Validation Dataset

The validation data set included in this shared task comprised 400 sentences, with 100 sentences from each dialect, and was used for both model tuning and evaluation purposes.

In some of the provided examples, there is a metaphorical expression that adds complexity to the translation process. For example, the phrase below employs imagery and cultural nuances that may not directly translate into other languages. It conveys a sense of agreement among the three individuals involved, using the metaphor of baking dough and leavening to suggest a collaborative effort. The metaphorical usage of words like "اللت" and "العجن الزيادة" adds depth and richness to the expression, but also presents a challenge for translators in capturing the essence of the message accurately in the target language.

{ "dialect": "Palestinian",
"source": "طالما ثلاثتكم متفقين لليش اللت و العجن الزيادة ،
خلينا نسحب حالنا و نطيح عالزلة و نعتذرله حقه علينا",
"target":
"بما أن ثلاثتكم متفقون، لماذا الخوض في الكلام الفارغ، دعونا
نذهب وننزل إلى الرجل ونعتذر له حقه علينا."
"English translation": "Since the three of you agree, why engage in nonsense, let's go down to the man and apologize to him for his right against us."}

3.1.2 Test Dataset

The test set for Subtask 3 comprises 2,000 sentences, with 500 sentences representing each of the four dialects: Egyptian, Emariti, Jordanian, and Palestinian. Numerous lengthy sentences pose further challenges in translation, illustrated by the below example for a sentence of 58 words:

{ "إنته بواب البناية صديق الناس صديق العوايل، عيب عليك،
يا أخي لين متى انته بس تشحت تشحت تشحت، عيب،
مرة وحدة في حياتك سو خير لوجه الله، ياخي ما عندكم شي
اسمه لوجه الله، ما عندكم لا نريد منكم جزاءً و لا شكورا،
ما عندك مصطلح طيب مرة
وحدة، عيب، بس بس مب زين" }

3.2 Fine-tuning Dataset

3.2.1 MADAR Dataset

We utilized the MADAR Arabic Dialect Corpus and Lexicon (Bouamor et al., 2018) in our study to fine-tune and Few-shot Fine-tuning models, serving as a crucial resource for machine translation research, particularly in handling Arabic dialects' complexities. This dataset includes 25 parallel translations, covering 2,000 sentences from 25 cities across the Arab world. In our experiments, we grouped the dataset into three categories, reflecting major geographical and dialectal regions. We focused on the cities of Salt, Amman, Cairo, Alexandria, and Jerusalem.

3.2.2 Custom Dataset

In our endeavor to create a dataset for fine-tuning, we conducted a series of "back translation" experiments leveraging Google Translate to generate translations between Arabic and English, as well as between Arabic and French. We translated Arabic sentences into English/French using Google Translate, subsequently back-translating the English/French translations into Arabic. This process aimed to introduce variation and diversity into the dataset, potentially capturing different nuances and interpretations of the original Arabic sentences. The resulting back-translated Arabic sentences served as the source (src) for the subsequent fine-tuning experiments. The rationale behind choosing French is twofold: French is widely spoken in several Arab countries, providing relevant dialectal nuances, and French translations can introduce further linguistic diversity, potentially capturing different structural and lexical variations that might not be present in English translations.

4 Methodology

4.1 Zero- and Few-Shot

We conduct a comprehensive evaluation of GPT3.5 (Ouyang et al., 2022), GPT4 (OpenAI, 2023), and Jais (G42, 2023) focusing on their ability to translate various dialects into modern standard Arabic. Specifically, we assess GPT3.5, GPT-4 in both zero-shot and few-shot scenarios. Following the findings of Kadaoui et al. (2023), which identify the few-shot setting as the most effective for a broad spectrum of Arabic to English translation tasks, we select two examples in this setting. Additional details about our experimental prompt are provided in Table 3 for GPT3.5 and GPT-4, and in Table 4 for Jais.

Zero-shot. We evaluate GPT models and Jais in a zero-shot setting with a simple prompt asking the model to translate dialectal Arabic into MSA. Specifically, for the models, we use a temperature of 0.1 and top_p value of 0.4 to ensure a stable and consistent text generation.

Few-Shot. We also use GPT3.5 and GPT-4 in the 2-shot setting and Jais in the 1-shot by providing examples from each dialect. We keep the example static throughout the dialect.

Combination approach. Furthermore, we noticed that Jais already has a pre-existing familiarity with Jordanian and Palestinian dialects, as evidenced by Table 1. Therefore, we implement a new strategy exclusively for Emirati and Egyptian dialects (i.e. One-shot for Egyptain and Emirati, and Zero-shot for Jordanian and Palestinian). This particular approach has yielded optimal outcomes across all methodologies.

4.2 Model Fine-tuning

4.2.1 NLLB

The NLLB model was subsequently fine-tuned using the custom dataset to evaluate whether initially providing it with a basic translation from Google Translate would enhance its performance by simply making corrections to this preliminary output. This approach aimed to determine whether starting with an approximate translation and refining it could lead to more accurate results than other methods.

4.2.2 Jais

We undertook the task of fine-tuning Jais, a pre-trained model, utilizing a custom dataset using a LoRa PEFT method. LoRa enables fine-tuning a small number of additional weights in the model while freezing most of the parameters in the pre-trained network. One of the main advantages of LoRa is that it preserves the original weights, mitigating the risk of catastrophic forgetting.

This attempt was pursued to evaluate Jais' efficacy, as it was exclusively trained on Arabic data. However, our findings revealed that Jais model exhibited superior performance without additional training, thereby showcasing its inherent robustness and adaptability across linguistic domains. The results are presented in Table 1.

5 Results

For our results' analysis, translations were assessed using two main criteria: (i) the performance of MT systems on each dialect separately, and (ii) the average score across all four dialects. This dual evaluation strategy enables a thorough evaluation of system performance, taking into account both dialect-specific intricacies and overall translation quality.

As we can see in Table 1, Jais, with a combination approach setup with Zero- and Few-shot, provides the highest 19.48 average BLEU score in our project. It means this model is already enhanced during the initial training process and provides optimal performance without fine-tuning. Additionally, we obtained results closest to our upper bound us-

Model	Egyptian	Emirati	Jordanian	Palestinian	Avg
Jais (combination approach)	17.14	22.22	16.82	21.56	**19.48**
Jais (0-shot)	16.33	19.48	17.05	21.61	**18.72**
Jais (1-shot)	17.14	22.22	6.77	9.04	13.83
Jais (custom data)	9.06	9.04	8.17	11.49	9.58
NLLB (custom data)	8.51	8.33	9.90	8.34	8.77
GPT-3.5 (0-shot)	13.39	10.92	15.10	15.80	14.06
GPT-3.5 (2-shot)	15.63	15.50	15.83	18.41	**16.67**
GPT-4 (0-shot)	14.58	14.28	19.00	17.17	16.30
GPT-4 (2-shot)	15.40	14.45	17.17	18.08	16.53
google translate	14.30	9.95	11.62	11.30	11.93

Table 1: BLEU score on the *val* dataset.

ing GPT versions 3.5 and 4 models - with 16.67 and 16.53 respectively.

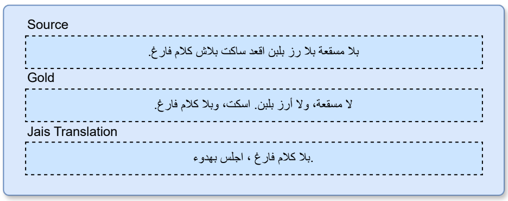
Model	Egyptian	Emirati	Jordanian	Palestinian	Avg
Jais (combination approach)	16.57	23.38	21.37	20.62	20.44

Table 2: BLEU score on the *test* dataset for the best setups.

6 Discussion

In machine translation (MT) across Arabic dialects and Modern Standard Arabic (MSA), our study within the NADI shared task's Subtask 3 illuminates the challenges and advancements in this domain. The final results for the test set (Table 2) show that a combination approach setup of Jais provides the highest 20.44 average BLEU score. Moreover, the selected GPT-3.5 model within a 2-shot approach obtained 0 for every BLEU score segment, indicating the need for further refinement or exploration of alternative methodologies. The divergence in results of GPT-3.5 model could potentially be attributed to the diverse nature of the test set and the propensity for the model to generate hallucinations, leading to discrepancies in BLEU scores. By evaluating models such as Jais, NLLB, GPT-3.5, and GPT-4, our findings underscore Jais' remarkable proficiency, surpassing all others in the combination approach setting. A significant factor contributing to Jais' superior performance is its specific pre-training for the Arabic language, which enhances its ability to handle various dialects effectively. In contrast, models like GPT-3.5 and GPT-4, while powerful, are general-purpose language models not specifically optimized for Arabic. This lack of specialization likely contributes to their lower performance in this specific task. Our study did not include other Arabic-specific models due to the focus on widely-used, general-purpose models to benchmark against a specialized model like Jais. Future work should consider incorporating additional Arabic-specific models to provide a more comprehensive evaluation and better understand the capabilities and limitations of different approaches in translating Arabic dialects.

Error Analysis. After examining samples of the translated source sentences, we observed that named entities or country-specific words, including dish names, were the primary sources of errors, as illustrated in Figure 1.



Source

بلا مسقعة بلا رز بلبن اقعد ساكت بلاش كلام فارغ.

Gold

لا مسقعة، ولا أرز بلبن. اسكت، وبلا كلام فارغ.

Jais Translation

بلا كلام فارغ ، اجلس بهدوء.

Figure 1: Translation Sample for Error Analysis.

7 Conclusion

Subtask 3 of the NADI shared task is a crucial avenue for advancing Arabic dialect machine translation, encouraging exploration and comparison of diverse translation techniques. By fostering the development of novel datasets and evaluating system efficacy across varied dialects, we've unearthed robust and universally applicable machine translation approaches, poised to enhance communication for Arabic speakers worldwide. Our findings underscore the profound influence of training data on model proficiency, elucidating the superior performance of models such as Jais in contrast to those not primarily tailored for Arabic, such as GPT and NLLB. As an additional achievement of our project, we took second place among all participants in the competition.

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. 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).

Wael Farhan, Bashar Talafha, Analle Abuammar, Ruba Jaikat, Mahmoud Al-Ayyoub, Ahmad Bisher Tarakji, and Anas Toma. 2020. Unsupervised dialectal neural machine translation. *Information Processing & Management*, 57(3):102181.

G42. 2023. Jais. https://inceptioniai.org/jais/. Accessed: [24/04/2024].

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.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. *arXiv preprint arXiv:2310.13661*.

OpenAI. 2023. GPT-4 Model. https://www.openai.com. Accessed: [24/04/2024].

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. *Preprint*, arXiv:2203.02155.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

A Appendix

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: []

Table 3: Zero-shot and few-shot templates. We formatted the prompt with appropriate input and examples before feeding it to GPT.

Shot	Prompt
Zero-shot	ترجمة نص من اللهجة العربية التالية إلى اللغة العربية الفصحى الحديثة: {dialect_text}:{dialect} اللغة العربية الفصحى الحديثة: []

Few-Shot	ترجمة نص من اللهجة العربية التالية إلى اللغة العربية الفصحى الحديثة:
	{dialect_text_2}:{dialect_2}
	اللغة العربية الفصحى الحديثة: {MSA}
	{dialect_text}:{dialect}
	اللغة العربية الفصحى الحديثة: []

Table 4: Zero- and few-shot templates using formatted prompt with appropriate input and examples before feeding it to Jais.