Alson at NADI 2024 Shared Task: Alson - A Fine-Tuned Model for Arabic Dialect Translation

Manan AlMusallam[1]
[1]Computer Science Department,
Imam Mohammad Ibn Saud Islamic University,
Riyadh, Saudi Arabia
mmmusallam@imamu.edu.sa and **Samar Ahmed**[2]
samar.sass6@gmail.com

Abstract

DA-MSA Machine Translation is a recent challenge due to the multitude of Arabic dialects and their variations. In this paper, we present our results within the context of Subtask 3 of the NADI-2024 Shared Task (Abdul-Mageed et al., 2024) that is DA-MSA Machine Translation. We utilized the DIALECTS-MSA MADAR corpus (Bouamor et al., 2018), the Emi-NADI corpus for the Emirati dialect (Khered et al., 2023), and we augmented the Palestinian and Jordanian datasets based on NADI 2021. Our approach involves developing sentence-level machine translations from Palestinian, Jordanian, Emirati, and Egyptian dialects to Modern Standard Arabic (MSA). To address this challenge, we fine-tuned models such as (Nagoudi et al., 2022) AraT5v2-msa-small, AraT5v2-msa-base, and (Elmadany et al., 2023) AraT5v2-base-1024 to compare their performance. Among these, the AraT5v2-base-1024 model achieved the best accuracy, with a BLEU score of 0.1650 on the development set and 0.1746 on the test set.

1 Introduction

The Arabic language includes two main variants: Modern Standard Arabic (MSA), which is a formal language used in newspapers, news broadcasts, and literature, and Dialects Arabic (DA), which contains a diverse array of informal languages spoken across Arab countries, each distinguished by its own unique characteristics (Almansor, 2018). DA-MSA Machine Translation is a relatively recent development in machine translation (MT), largely due to the spread of social media and the widespread use of dialects in daily interactions and online communication. Given the multitude of Arabic dialects and their variations. Creating accurate tools to process Arabic social media content presents significant challenges (Derouich et al., 2023). Therefore, there is a need to unify these dialects under the umbrella of MSA, which serves as a standard form of Arabic. DA-MSA translation not only facilitates communication between speakers of different Arabic dialects but also enhances the comprehension of written or spoken content for those who primarily use Standard Arabic. Despite its importance, the availability of datasets for DA-MSA translation is limited compared to English datasets. Therefore, translation in English has preceded Arabic. In our research, we focus on translating four Arabic dialects (Egyptian, Emirati, Jordanian, and Palestinian) into MSA using the MADAR parallel dataset and Emi-NADI for Emirati dialect. We augmented this dataset by extracting data from the NADI2021 dataset for Jordanian and Palestinian dialects. Additionally, we utilized ChatGPT to translate these dialects into MSA, demonstrating the capabilities of large language models. Throughout our experiments, we fine-tuned model's such as AraT5v2-msa-small, AraT5v2-msa-base AraT5v2-base-1024 to achieve accurate MSA translations. This paper is organized as follows: In Section 2, we review related work. In Section 3, we present our methodology and dataset. In Section 4, we evaluate the models accuracy and present the results. In Section 5, we provide a discussion of the results. Finally, we conclude with an analysis of errors observed during the experimentation.

2 Related Work

Some researchers have concentrated on translating a single dialect into Modern Standard Arabic. For instance, the architecture of an RNN sequence-to-sequence learning model (Encoder-Decoder Model) was adopted in (Al-Ibrahim and Duwairi, 2020) to translate the Jordanian Dialect to MSA. The researchers trained this model to accurately learn the translation probability of Jordanian words/phrases to their corresponding standard counterparts, manually collecting the dataset for training. In another study, the focus was on the

Proceedings of The Second Arabic Natural Language Processing Conference, pages 764–768

Tunisian Dialect (Sghaier and Zrigui, 2020). The objective was to develop a system that translates text written in Tunisian Dialect (TD) into Modern Standard Arabic (MSA) using a rule-based approach. This study (Faheem et al., 2024) also focused on one dialect translating Egyptian dialects to Modern Standard Arabic (MSA) by combining supervised and unsupervised learning techniques in neural machine translation to improve accuracy and reliability. On the other hand, efforts have been directed towards translating multiple dialects. (Derouich et al., 2023) focused on machine translation of Arabic dialects to Modern Standard Arabic (MSA) using the DIALECT-MSA MADAR corpus. The researchers employed pre-training of the AraT5 transformer model and compared its performance with existing transformer models. They also utilized the back-translation method, which involves using English as an intermediary language between Dialectal Arabic and MSA, leading to improved translation quality. They achieved a BLEU score of $11.14\,\%$ on the dev set and $10.02\,\%$ on the test set. In another study, (Khered et al., 2023) the emphasis was on machine translation of text from four Arabic dialects to Modern Standard Arabic. The researchers fine-tuned three types of T5 models, namely, AraT5, mT5, and mT0. Additionally, they developed a new dataset, Emi-NADI, which contains MSA translations of sentences written in Emirati. In our work, we emulate the second approach and focus on translating multiple Arabic dialects into Modern Standard Arabic.

3 Methodology

3.1 Dataset

This section describes the datasets utilized in training our models, as shown in Table 1. Our approach began with the use of the MADAR parallel corpus, which includes data covering the dialects of 25 Arabic-speaking cities, as well as English and Modern Standard Arabic (MSA). It is important to note that the MADAR corpus does not include a parallel corpus for the Emirati dialect. To address this gap, we incorporated another parallel corpus called Emi-NADI. Additionally, we augmented the dataset for Jordanian and Palestinian dialects, which were underrepresented in the MADAR corpus. We extracted data from the NADI2021 dataset for these dialects, yielding 420 instances for the Palestinian dialect and 426 instances for the Jordanian dialect. To showcase the capabilities of large language models, we employed OpenAI's ChatGPT, specifically the GPT-3.5 version, to translate Jordanian and Palestinian dialects into MSA. Our approach involved crafting prompts for ChatGPT to carry out these translations. This process was facilitated through the web user interface provided by OpenAI, enabling us to interact directly with ChatGPT via a browser. The model was used in its original form without any additional training or fine-tuning. After generating the translations, we thoroughly reviewed the content to assess its quality and accuracy. Finally, we performed preprocessing on the dataset extracted from NADI2021. This preprocessing included removing URLs, punctuation marks, digits, extra white spaces, and vowels to clean and standardize the data for training.

Dataset	**Total**
MADAR + Jor-Pla- NADI	91,758
4 dialects + Jor-Pla-NAD	20, 557
Emi-NADI	2,758
Jor-Pla- NADI	846

Table 1: Dataset

NADI provided both development and testing datasets. The development data comprised 400 entries, with 100 entries for each dialect. Each entry in the development data had a corresponding MSA equivalent. In contrast, the testing data consisted of 2000 entries, with 500 entries per dialect, but did not include MSA equivalents.

3.2 Fine-tuning pre-trained Language Models

We utilized the AraT5v2 model, Arabic Text-To-Text Transfer Transformer, which was trained on an extensive corpus of Arabic text (both unlabeled and labeled) that includes all categories of Arabic (i.e., Classical Arabic (CA), Dialectal Arabic (DA), and Modern Standard Arabic (MSA)). The model was fine-tuned for specific downstream tasks, including translation, summarization, question answering, and text generation (Elmadany et al., 2023). In our experiments, we used three variants from the AraT5v2 series: AraT5v2-msa-small, AraT5v2-msa-base, and AraT5v2-base-1024. We chose AraT5v2-base-1024 as the basis for our experiments due to its high accuracy.

3.3 Hyperparameter Optimization

We trained our models using Nvidia A100 GPUs in Google Colab. All models accepted input se-

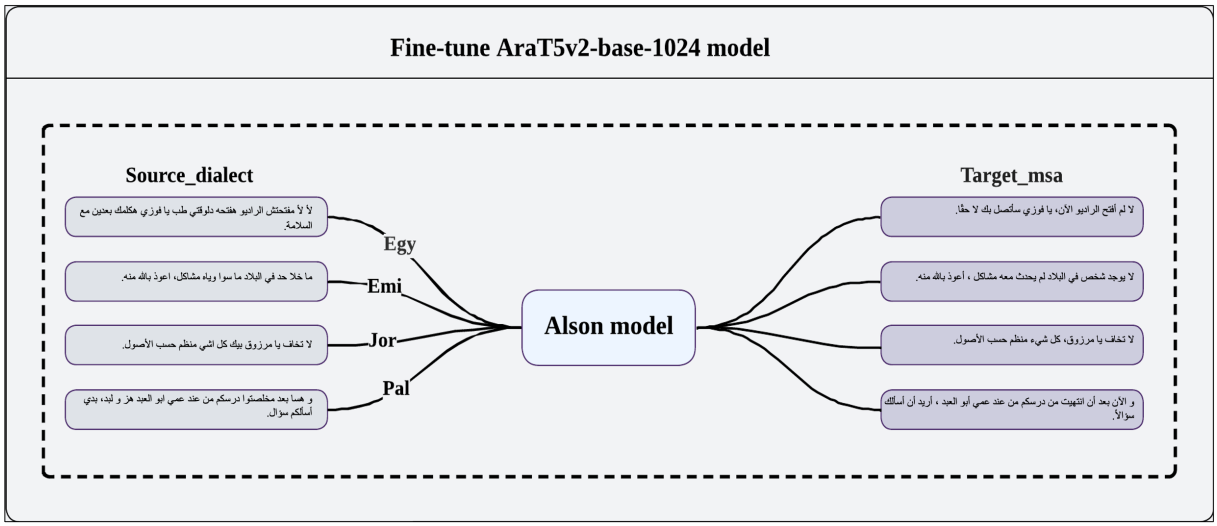

Figure 1: An example of alson model

quences with a maximum length of either 105 or 128 tokens and generated output text with the same respective maximum lengths of either 105 or 128 tokens. The learning rates were configured to either 5e-5 or 7e-5, while batch sizes were set to 16 or 8, chosen based on the model's complexity and resource requirements. The maximum number of epochs was set to 20.

4 Results

All models for Subtask 3 were evaluated using the Bilingual Evaluation Understudy (BLEU), a metric for automatically machine translated text (Papineni et al., 2002). We obtained two sets of results: one set for the 4 dialects + Jor-Pla-NADI dataset, and another set for the MADAR + Jor-Pla-NADI dataset. The results for the 4 dialects + Jor-Pla-NADI dataset outperformed those for the MADAR + Jor-Pla-NADI dataset in terms of accuracy. We utilized three variants from the AraT5v2 series: AraT5v2-msa-small, AraT5v2-msa-base, and AraT5v2-base-1024. Our initial experiments were conducted using AraT5v2-msa-small, which produced low accuracy, with the highest achieved accuracy being 0.024 at epoch 30. Subsequently, we employed AraT5v2-msa-base, which although better and more complex than the first model, still resulted in low accuracy, albeit slightly improved with the highest achieved accuracy of 0.056 at epoch 10. Finally, we proceeded with AraT5v2-base-1024 as the last variant, which proved to be the most effective among them. The results were based on experiments conducted on two datasets with different epochs. We observed that in all experiments, the 4 dialects + Jor-Pla-NADI dataset consistently outperformed the MADAR + Jor-Pla-NADI dataset across randomly different epochs. This comparison was conducted to determine which dataset yielded better performance, as illustrated in Figure 2. And the highest result we achieved on the development dataset, trained on the 4 dialects + Jor-Pla-NADI dataset, was 0.1650. Additionally, the highest result we achieved on the test dataset, also trained on the 4 dialects + Jor-Pla-NADI dataset, was 0.1746. These results were obtained with a learning rate of 7e-5, a batch size of 16, and an epoch of 2. See Table 2 for details. Figure 1 shows sample output generated by the alson model, randomly selected from the prediction file. These examples illustrate the model's performance, demonstrating that it achieves satisfactory results.

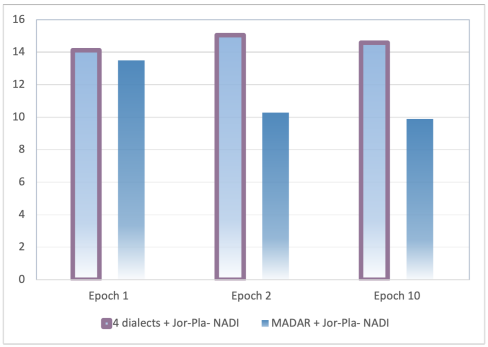

Figure 2: Model accuracy

Model	Overall	Egy.	Emi.	Jor.	Pal.
			AraT5v2-base-1024		
Dev	0.1650	0.1709	0.1273	0.1817	0.1618
Test	0.1746	0.1676	0.1253	0.2094	0.1843

Table 2: Model Performance on Development and Test Sets

5 Discussion

In our study, we observed that the AraT5v2-base-1024 model, which was trained on four dialects + Jor-Pla-NADI, outperforms other models. This is due to its training on a large and diverse Arabic dataset. This suggests that pre-trained models on extensive and diverse Arabic language data achieve better performance in translating dialects to Modern Standard Arabic (MSA). We also noted that the dataset with the four dialects + Jor-Pla-NADI yields better results than the MADAR + Jor-Pla-NADI dataset because focusing on a specific number of dialects helps the model learn more effectively. Conversely, training on a large number of dialects may weaken the model's predictive ability and overall performance. Additionally, the similarity between dialects and the variation in word meanings across dialects may additionally compromise the model's effectiveness. On adjusting the model parameters, such as batch size and max target input, we experimented with values but noticed only marginal performance differences. Similarly, varying the learning rate yielded high accuracy with values of 5e-5 and 7e-5. Increasing the epoch count to 20 or 15 did not significantly impact accuracy but consumed more time and resources. In fact, it sometimes even resulted in a decrease in accuracy. Furthermore, our model achieved low BLEU scores when trained on AraT5v2-msa-small and AraT5v2-msa-base models. At the level of each dialect separately, we noticed that the Egyptian (Egy) dialect, despite having a high number of rows in the training data, did not achieve highest accuracy. This could be due to the influence of Modern Standard Arabic (MSA). On the other hand, the Emirati (Emi) dialect showed lowest accuracy, possibly due to foreign dialects. The Jordanian (Jor) and Palestinian (Pal) dialects achieved high accuracy, with minimal differences between them, likely because of their high similarity to each other and to MSA. This similarity enhances the model's performance in translating these dialects to MSA.

6 Error Analysis

In our research, we conducted a comprehensive error analysis of our language model, which was specifically trained to translate four distinct Arabic dialects Egyptian, Emirati, Jordanian, and Palestinian into MSA. Following our error analysis, we observed that the model faces considerable challenges with dialect-specific words and culturally nuanced expressions, particularly in the Egyptian and Emirati dialects. For instance, in the Egyptian dialect, the letter (ج) is often pronounced as (ق) and in the Emirati dialect, (ك) is frequently altered to (ج). These phonetic changes are unique to each dialect and often lead to mistranslations or the model failing to translate the words. This difficulty arises because these words far significantly from their MSA counterparts. Consequently, our findings suggest that the closer a word or phrase is to MSA, the more effectively the model can translate it. Conversely, as the words become more distinct from MSA, the model's ability to translate them declines. Furthermore, the model sometimes fails to translate verbs into their correct tenses or accurately translate repetitive structures in dialects. This indicates a need for the model to be more finely attuned to the subtleties and nuances of each dialect. Fine-tuning with specialized datasets that capture these dialect-specific characteristics is essential for enhancing the model's translation accuracy. Improving the model's ability to manage these phonetic and structural variations is crucial for better performance. As illustrated in Figure 3, the current model output shows several instances of poor translation for each dialect.

Dialects	Source_da	Target_msa
Egyptian	جنازة حسام هتتعمل لما نتطمن على بابا وعزاه بتأجل لحد ما ناخد بتاره ولا حضرتك ناسي اننا صعايدة.	جنازة حسام ستتعمل عندما نتطمن على بابا وعزاه، ولا تنسى أننا صعايدة.
Emirati	طفرتينا، أذيتينا براقعيه وبراقعيه، عن عيد العذ، نأكل براقع نلبس براقع، الله يهديج. هيا بنا نطرشلج براقع، الله يهديج.	طفرتينا، أذيتينا براقعيه وبراقعيه، عن عيد الغد، نأكل براقع نلبس براقع، الله يهديج. هيا بنا نطرشلج براقع، الله يهديج.
Jordanian	لقد أزهقت من الموضوع أزهقت تماما لم أعد قادرا علي التحمل	لقد أزهقت من الموضوع أزهقت تماما لم أعد قادرا علي التحمل
Palestinian	والله وحياة كل شعرة فراس فاطمة مني مفلتهن غير فالمطبخ	والله وحياة كل شعرة فراس فاطمة مني مفلتهن غير في المطبخ

Figure 3: Translation Examples for Different Dialects to MSA

7 Conclusion, Future Work, and limitations

In this work, we introduced our fine-tuned AraT5v2-base-1024 model for NADI Subtask 3 (Open Dialect-to-MSA), aimed at translating four

dialects to MSA. While our results are satisfactory, there's always opportunity for improvement. In future directions, we plan to expand our translation capabilities to include more Arabic dialects using APIs such as GPT or other large language models designed specifically for the Arabic language. Additionally, we aim to engage volunteer reviewers to assess the quality of translations. We also envision practical applications for our work in the real-world. Regarding limitations, we faced resource constraints on Google Colab Pro, which prevented us from conducting more experiments with larger datasets. We encountered issues with RAM limitations, causing occasional crashes during model testing. To address this, we divided the test file into ten segments, which were then concatenated and uploaded to CodaLab. However, we recognize the need for larger servers with advanced resources to support our tasks effectively.

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)*.

Roqayah Al-Ibrahim and Rehab M Duwairi. 2020. Neural machine translation from jordanian dialect to modern standard arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 173–178. IEEE.

Ebtesam Hussain Almansor. 2018. *Translating Arabic as low resource language using distribution representation and neural machine translation models*. Ph.D. thesis.

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).

Wiem Derouich, Sameh Kchaou, and Rahma Boujelbane. 2023. ANLP-RG at NADI 2023 shared task: Machine translation of Arabic dialects: A comparative study of transformer models. In *Proceedings of ArabicNLP 2023*, pages 683–689, Singapore (Hybrid). Association for Computational Linguistics.

AbdelRahim Elmadany, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Octopus: A multitask model and toolkit for Arabic natural language generation. In *Proceedings of ArabicNLP 2023*, pages 232–243, Singapore (Hybrid). Association for Computational Linguistics.

Mohamed Atta Faheem, Khaled Tawfik Wassif, Hanaa Bayomi, and Sherif Mahdy Abdou. 2024. Improving neural machine translation for low resource languages through non-parallel corpora: a case study of egyptian dialect to modern standard arabic translation. *Scientific Reports*, 14(1):2265.

Abdullah Khered, Ingy Abdelhalim, Nadine Abdelhalim, Ahmed Soliman, and Riza Batista-Navarro. 2023. UniManc at NADI 2023 shared task: A comparison of various t5-based models for translating Arabic dialectical text to Modern Standard Arabic. In *Proceedings of ArabicNLP 2023*, pages 658–664, Singapore (Hybrid). Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Mohamed Ali Sghaier and Mounir Zrigui. 2020. Rule-based machine translation from tunisian dialect to modern standard arabic. *Procedia Computer Science*, 176:310–319.