ANLP-RG at NADI 2023 shared task: Machine Translation of Arabic Dialects: A Comparative Study of Transformer Models

Wiem Derouich and **Sameh Kchaou** and **Rahma Boujelbane**
ANLP Research Group, MIRACL Lab. FSEGS,
University of Sfax, Tunisia
wiemderwich123@gmail.com samehkchaou4@gmail.com rahma.boujelbane@gmail.com

Abstract

In this paper, we present our findings within the context of Subtask 2 of the NADI-2023 Shared Task. This task requires the exclusive utilization of the DIALECT-MSA MADAR Bouamor et al. (2018) corpus to develop sentence-level machine translations from Palestinian, Jordanian, Emirati, and Egyptian dialects to Modern Standard Arabic (MSA). However, MADAR lacks a parallel Emirati-MSA corpus. To address this challenge, we pre-trained the AraT5 transformer model using different configurations of the MADAR corpus and compared their performance results with those of existing transformer models. The best model achieved a BLEU score of 11.14% on the dev set and 10.02% on the test set.

1 Introduction

Arabic dialects (AD) represent a diverse range of informal languages spoken throughout Arab countries. The rise of social media has greatly amplified the growth of these dialects, which have become an integral part of everyday communication. Platforms such as Twitter, Facebook, and Instagram often feature user-generated content written in these dialects. Arabic dialects can indeed vary significantly from one region to another, and these vast regional differences make them challenging to understand and interpret. This linguistic diversity can be so pronounced that even within a single country, identical words might bear different meanings. As a result, due to the variation among these Arabic dialects (ADs), it becomes exceedingly challenging to create tools capable of accurately processing Arabic social media content. It can also be difficult to employ standard tools designed for Modern Standard Arabic (MSA), which serves as the mother language for these dialects. One solution to overcome this problem involves leveraging the richness of the MSA language by translating the dialect variants into it. Currently, there is a significant amount of work focused on translating dialects to MSA. However, most approaches treat each dialect separately, as seen in studies like Kchaou et al. (2022) for Tunisian and Al-Ibrahim and Duwairi (2020) for the Jordan dialect. Yet, it's important to acknowledge that these dialectal variations coexist in social networks. Therefore, it is important to develop models capable of handling the processing of all these dialects collectively. This work fits into this context by involving the development of a machine translation (MT) model to translate a subset of Arabic dialects, namely Palestinian and Levantine, into Modern Standard Arabic (MSA). As part of the competition offered by the NADI shared task, the challenge involves the development of a translation model for four dialects based solely on the MADAR corpus. It's worth noting that MADAR lacks a parallel corpus for the Emirati dialect. In this paper, we outline the experiments to build a dialect translation model. Specifically, we have compared the results of two methods: the first involves fine-tuning the AraT5 transformer model (Nagoudi et al., 2021) utilizing various corpus configurations from MADAR, while the second entails refining existing tools and employing the back-translation method. The rest of this paper is structured as follows: Section 2 outlines related works. Section 3 describes the dataset used. The fine-tuning of AraT5 models is presented in Section 4. We assess the benefits of leveraging tools to improve the translation process in Section 5. In Section 6, we discuss the results obtained. Finally, Section 7 provides a conclusive summary.

2 Related works

In the field of neural machine translation for Arabic dialects, the primary focus has been on translating these dialects into Modern Standard Arabic (MSA). However, most of these works typically concentrate on a single dialect, resulting in a lack of models that address the full spectrum of dialects. For instance, Al-Ibrahim and Duwairi (2020) conducted

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 683–689
December 7, 2023 ©2023 Association for Computational Linguistics

a study that focused on translating the Jordanian Arabic dialect into MSA using deep learning techniques and implemented an RNN encoder-decoder model. However, the size of the corpus limited their progress. Similarly, Baniata et al. (2018) tackled the challenge of translating Levantine dialects, including Jordanian, Syrian, and Palestinian, into MSA. They worked with a relatively small dataset of about 20,000 parallel sentences from the AD Applications and Resources (MADAR) and Parallel Arabic Dialect Corpus (PADIC) corpora. Their approach introduced an RNN-based multitask learning model in which the decoder was shared across language pairs, with each source language having its own encoder. A transductive transfer learning approach, introduced by Hamed et al. (2022), emerged in the context of low-resource neural machine translation for the Algerian Arabic dialect. This approach employed fine-tuned transfer learning to transfer knowledge from a parent model to a child model. The evaluation was carried out using the MADAR and PADIC corpus. This study applied the transductive transfer learning strategy with two NMT models: Seq2Seq and Attentional-Seq2Seq. Moreover, Nagoudi et al. (2022) introduced TURJUMAN, a versatile neural toolbox that can translate 20 languages into MSA. The TURJUMAN toolbox uses the power of the AraT5 model, renowned for its ability to decode Arabic. Notably, TURJUMAN allowed for flexibility in decoding methods, facilitating the creation of paraphrases for MSA translations. The tool was trained to use semantic similarity to collect publicly available parallel data samples to ensure data quality. This initiative resulted in the development and launch of AraOPUS-20, which establishes a new benchmark for machine translation. It encompasses a benchmark dataset (AraOPUS-20) and the translation toolkit (TURJUMAN). Another contribution comes from Kchaou et al. (2023), who introduced a hybrid approach to building a translation model for the Tunisian dialect. They proposed different augmentation methods to create a large corpus. Using this corpus, the authors tested different NMT models. The best model was obtained using JoeyNMT, achieving a BLEU score of 69.22

3 NADI-2023 Shared Task Subtask 2: DATASETS

In Subtask 2, we had access to three primary datasets: training (Train), development (Dev), and testing (Test), as outlined in Table 1. Our approach began with the utilization of the MADAR parallel corpus as our training set. During this phase, our model learned from the data and fine-tuned its parameters accordingly. Following the training, we evaluated our model's performance on the development set provided by the shared task. Finally, we generated translations using the best configuration on the test set. The subsequent sections will provide a detailed description of the contents within these three corpora.

Data set	**#Lines**
Train	111096
Dev	400
Test	2000

Table 1: Distribution of Different Sets.

3.1 Training SET

Subtask 2 allowed the usage of only the dataset from the MADAR parallel corpus for training. The statistics for the MADAR corpus for Subtask 2 are provided in Table 2.

Corpus	**#lines**	**#token**	**#vocabularies**
Tunisian	14k	87113	17102
Iraq	4k	2414	6466
Libya	4k	26209	6247
Morroco	14k	94289	18120
Syria	4k	6098	24363
SAUDI-ARABIA	4K	24751	6248
EGYPT	4k	26757	6239
JORDAN	42k	26074	6247
PALESTINIAN	2k	12574	3902
QATAR	12k	72878	12480
Yemen	2k	12823	4317
Algeria	2k	13198	4180
Lebanon	12k	72806	15531
Oman	2k	13201	4531
Sudan	2k	13352	4120

Table 2: Statistics of MADAR Subtask 2 Data Set.

3.2 Dev set

The development set comprises 400 sentences, with 100 sentences dedicated to each dialect. This dataset plays a crucial role in enhancing and evaluating translation systems, aiming for exceptional results. Each development tweet is accompanied by a unique identifier (#1_id) for each dialect, followed by the tweet's content (#2_content). The third column (#3_label) presents the tweet's gold label at the country-of-origin level.

3.3 Test set

The test set includes a total of 2,000 sentences, with an equal distribution across four different dialects: Egyptian, Emirati, Jordanian, and Palestinian. These tests have been thoughtfully designed to assess the capability of translation systems to effectively convert AD into MSA. Furthermore, each test tweet is accompanied by a unique identifier (#1_id) and specifies the dialect's name at the country-of-origin level (#2_dialect_id). [1]

4 Fine-Tuning AraT5 models

In light of the impressive performance showcased by the transformer architecture in Neural Machine Translation (NMT) of Arabic Dialects, as highlighted by Kchaou et al. (2023) and Nagoudi et al. (2022), our proposed strategy is to further harness this potential. Specifically, we intend to fine-tune the transformer AraT5 model using the MADAR corpus. This fine-tuning process is geared towards developing a specialized Machine Translation (MT) model capable of effectively handling the four dialects introduced for testing.

4.1 Architectures

In order to determine the most suitable AraT5 configuration for this task, we conducted fine-tuning on seven different architectures, including:

- The **AraT5 base** model by Abdul-Mageed et al. (2021): This model represents a modification of the T5 (Text-To-Text Transfer Transformer), finely tuned for the processing of Arabic text. It functions as a foundational model for various natural language processing tasks, encompassing text classification, text generation, and machine translation (MT). AraT5-base capitalizes on the Transformer architecture and pre-trained embeddings to effectively comprehend and generate Arabic text.

- The **AraT5v2-base-1024** model represents an enhanced iteration of AraT5-Base. In this version, the sequence length has been extended from 512 to 1024, denoted by the "1024". This expanded sequence length significantly augments the model's adaptability across various Natural Language Processing (NLP) tasks. Notably, the fine-tuning process of AraT5v2-base-1024 exhibits approximately 10 times faster convergence compared to its predecessor, AraT5-base. This accelerated convergence holds the potential to significantly expedite both the training and fine-tuning procedures, thereby enhancing overall efficiency. The selection of this model for our experiments was motivated by its exceptional performance, as demonstrated in Table 4, where it outperformed other models under the AraT5v2-Base category.

- The **Sultan-ArabicT5** model Alrowili and Shanker (2022) : It is another variant of the T5 model tailored for Arabic text processing. Similar to other T5-based models, Sultan-ArabicT5 is versatile and can be fine-tuned for a range of natural language processing (NLP) tasks. Specific features and details of this model may vary depending on the creator's objectives and training data.

- **AraT5-MSA-Small and AraT5-MSA-Base** Models Nagoudi et al. (2021): We evaluated two additional versions of the AraT5 model in our experiments: AraT5-MSA-Base and AraT5-MSA-Small, each tailored to meet specific requirements. The AraT5-MSA-Base is an upgraded AraT5 version that is well-equipped to handle a wide array of standard Arabic Natural Language Processing (NLP) tasks. It boasts a larger architecture and an increased number of parameters, making it particularly adept at intricate tasks that demand a deep understanding of the language. AraT5-MSA-Base is an excellent choice for research projects and applications that necessitate advanced linguistic modeling. AraT5-MSA-Small in contrast, is a streamlined iteration of the AraT5 model, optimized for efficient processing of MSA data. It operates at a faster pace and demands fewer computational resources compared to the "Base" version. This version is typically employed in applications where efficiency is a priority, without a significant loss in quality. The key distinction between these two models lies in their size and their suitability for various standard Arabic NLP tasks. AraT5-MSA-Small prioritizes speed and resource efficiency, while AraT5-MSA-Base excels in proficiency and

[1] https://github.com/Wiemder/Levantin-Dataset

versatility across a broader spectrum of standard Arabic NLP tasks.

- **AraT5-Tweet-Small and AraT5-Tweet-Base** Models: as presented by Nagoudi et al. (2021), AraT5-Tweet-Small and AraT5-Tweet-Base are specialized models meticulously crafted to tackle the unique linguistic challenges presented by social media content, particularly tweets and informal online discourse. These models are fine-tuned to specifically address the subtleties involved in translating Arabic dialects commonly found in user-generated content on platforms like Twitter. Their incorporation into our experiments equips us with the tools needed to effectively navigate the complexities associated with translating such content.

In specific scenarios, the transformer-based model "AraT5v2-base-1024" can indeed prove to be a valuable asset for traditional machine learning models. In our specific context, the proposed fine-tuning of AraT5 models offers several advantages. These pre-trained models can be further customized and optimized for specific Natural Language Processing (NLP) tasks, subsequently serving as input features or foundational models for various tasks within traditional machine learning. Transformer-based models, including AraT5, bring advanced capabilities for text preprocessing, encompassing tasks such as tokenization, embedding, and attention mechanisms. These preprocessing steps can be seamlessly integrated with traditional machine learning models that might lack such built-in capabilities. The fusion of predictions from a transformer-based model and a traditional machine learning model, often referred to as ensemble learning, frequently results in enhanced prediction accuracy. This is particularly valuable for tasks that necessitate the handling of both textual and structured data, creating a synergy that can lead to improved performance across a wide range of applications.

4.2 MADAR Configurations for AraT5 Fine-tuning

Our approach involved fine-tuning the aforementioned models with various configurations of the MADAR corpus. Initially, we conducted experiments using the entire corpus, and subsequently, we suggested the utilization of a subset of dialects from Palestine, Jordan, and Egypt. In a second phase, we incorporated dialects from geographically adjacent regions, namely Qatari and Saudi-Arabian dialects. All the models used in this research were sourced from the Hugging Face repository, and the experiments were designed and executed using the PyTorch Transformers library. To ensure consistency and comparability, we implemented the models with identical parameter settings, as outlined in Table 3. This standardized approach enabled us to make meaningful comparisons and draw reliable conclusions from our experiments. These parameters were carefully selected to achieve optimal performance while minimizing training time. They were carefully selected to achieve optimal performance while minimizing training time. The maximum length for the number is set at 128 characters, and the batch size parameter is configured for training with a value of 16. We carried out a single training epoch to compare the initial performance of the model across various experiments. The sequence length of 20 characters was determined based on the improvement of the results. A learning rate of 2e-5 was optimal to achieve fast convergence without the risk of overfitting. The weight decay is sustained at 0.01 to regulate model learning, and a save_total_limit of 3 is used to retain essential checkpoints during training. These parameters are pivotal in ensuring the reproducibility of our experiments and the accuracy of our results. Table 4 provides a comprehensive view of the BLEU scores obtained for diverse Arabic dialects (ADs) generated by a range of models and strategies. Notably, the MADAR corpus, in combination with the AraT5v2-base-1024 model, emerges as the top performer with an impressive overall BLEU score of 11.14. This underscores the critical importance of meticulous model selection in achieving optimal translation quality for specific Arabic dialects. Additionally, the variability in BLEU scores across different dialects suggests that certain models may exhibit superior performance for specific dialects, reinforcing the need for tailored approaches to enhance translation quality effectively.

5 Leveraging existing tools

To elevate the BLEU scores in our translation task, we pursued enhancements through the utilization of existing tools. Our approach unfolded in two key steps: Firstly, we showcased the efficacy of these tools in translating dialects into Modern Standard Arabic (MSA). Secondly, taking advantage of the

Parameters	Values
Max-length	128 Characters
Batch-size	16
Epoch	1
Seq-length	20
Learning-rate	2e5
Weight-decay	0.01
Sav-total-limit	3

Table 3: Parameters of the AraT5v2-base-1024 model

Corpus	Model	Overall	Egy	Emirate	Jord	Pales
TS MADAR	AraT5v2-base	11.14	10.58	8.11	10.04	11.38
TS MADAR	Sultan-ArabicT5	6.11	5.03	6.46	5.69	6.80
Egy-Pal-Jor	AraT5V2-Base	5.62	4.56	5.46	6.19	5.58
Egy-Pal-Jor-Qat-Ksa	AraT5V2-1024	7.02	6.51	6.16	8	6.38
Aug-ALE	AraT5V2-Base	9.52	10.66	6.88	8.76	8.95
Aug-ALE-ALX-JER	AraT5V2-Base	9.40	10.66	6.88	8.76	8.95
Aug-PaysGolf	AraT5V2-Base	6.09	5.40	7.28	4.88	5.88

Table 4: Bleu scores on the Dev set of the proposed configurations.

broader availability of Dialectal Arabic (DA) to English translation tools, we introduced the back translation method. This technique involves using English as an intermediary language between Dialectal Arabic and MSA, contributing to improved translation quality.

5.1 Direct-Translation

In our research, we leveraged the capabilities of TURJUMAN, a robust neural machine translation system designed not only for Modern Standard Arabic (MSA) but also for 20 other languages[2]. To optimize its performance, we carefully fine-tuned TURJUMAN with unique configurations. These included setting "search_method" to "beam," "seq_length" to 20, "num_beams" to 5, "no_repeat_ngram_size" to 3, and "max_outputs" to 1. These distinctive parameter choices allowed us to generate a fresh batch of MSA texts, resulting in a substantial improvement in BLEU scores, as depicted in table 5. Furthermore, we conducted experiments to explore the impact of increasing the value of max_outputs" to 3, thereby generating three distinct MSA texts. Remarkably, these experiments revealed no significant variation in BLEU scores among the different texts. Additionally, we experimented with the dl-translate 0.3.0 library[3], designed for text translation. Unfortunately, our evaluation using BLEU scores indicated that the quality of the generated texts fell below our expectations. These findings underscore the critical importance of tool selection and configuration in optimizing translation quality and ultimately enhancing BLEU scores in our research.

	Models	Overall	Egypt	Emirati	Jordan	Palestinian
Back translation	GoogleTranslator	10.98	12.12	7.73	9.54	12.18
	PonsTranslator	10.87	12.15	7.72	9.69	11.71
Direct-Translation	TURJUMAN	10.08	11.18	12.99	7.98	8.50

Table 5: Bleu Scores on the Dev set using existing transformer-based tools.

5.2 Back-Translation

In implementing the back-translation technique, following the approach by Hoang et al. (2018), we employed English as the intermediary language. This method was executed using the deep-translator library, which incorporates both the PonsTranslator and Google-translator models. As demonstrated in Table 5, this approach led to improvements in BLEU scores. Additionally, we re-utilized the dL-Translate 0.3.0 library for back-translation. This process entails the generation of English texts, followed by the subsequent back-translation into MSA. We applied this method to both transformer models offered in dl-translate 0.3.0: "nllb-200" and "m2m100." These efforts contributed to the enhancement of our translation quality and the corresponding BLEU scores, demonstrating the effectiveness of back-translation techniques in our research.

6 Discussion

The results we obtained exhibit a range of variations, encompassing both positive and negative outcomes. Our performance curve for the implemented strategies demonstrates fluctuations, highlighting the complexity of the translation task. It's important to note that the table of BLEU scores does not include certain results, such as those for AraT5-tweet-base, which received a NADI BLEU score of 0 and an overall BLEU score of 0.28. In contrast, the highest BLEU score of 11.14 on the dev set was achieved through the fine-tuning of AraT5V2-base-1024. Tis con

Moreover, upon analyzing the use of DeepL Translate and dl-translate 0.3.0, we observed that the models from the DeepL Translate library outperformed those from the dl-translate library. These models exhibited the potential to contribute to enhancing our corpus, resulting in higher MSA scores. In contrast, the results obtained from

[2] https://github.com/UBC-NLP/turjuman

[3] https://github.com/xhluca/dl-translate

the dl-translate library were significantly lower. Moreover, our experiments with data augmentation yielded limited benefits, potentially due to the extensive diversity of Arabic written forms. Table 6 provides examples generated using the augmentation method, offering insights into the outcomes of this approach. The relatively lower success rates observed in our experiments can be attributed to several factors. These include limitations within the corpus, notably the absence of the Emirati dialect in the parallel corpus. Additionally, the vocabulary and the quantity of comments available within the MADAR dataset may have also played a role in influencing the results. These factors collectively contribute to the challenges associated with achieving higher success rates in dialect translation tasks. Addressing these limitations and enhancing the availability of diverse and comprehensive datasets could potentially lead to improved translation outcomes in the future.

ALX-MADAR-Corpus	Aug1-ALX	Aug2-ALX
مّكن بدف الِك مية الي بميتين دولَار Can my $200 check be cashed?	دَا مّكن تكدر تدفء بقية اليك الِ بميتين ٥٠٠ دولَار dear Can my $500 check be 200 cashed	مّكن بس الي بَالّون اهمر الي بميتين دولَار please dear Can my $200 check red be cashed !

Table 6: Example of generated sentence using augmentation method.

7 Conclusion

Our work constitutes an integral contribution to the NADI2023 shared task, which centers around the machine translation of Arabic dialects (AD) into Modern Standard Arabic (MSA) using the MADAR corpus. Throughout our research, we explored a multitude of methods and strategies aimed at tackling this complex and challenging task. Our efforts yielded two notable strengths. Firstly, fine-tuning models, with a particular focus on the "AraT5v2-base-1024" model, emerged as an effective approach for enhancing translation quality. Additionally, we achieved commendable results by leveraging existing translation tools, especially the Google Translator model, coupled with back-translation methods. These outcomes underscore the relevance and practicality of these approaches for translating Arabic dialects. In fact, these results open up the possibility of utilizing these methods to automate the parallel corpus construction process. Furthermore, we are dedicated to furthering our research efforts by delving into additional fine-tuning techniques for transformer models.

References

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

R. Al-Ibrahim and R. M. Duwairi. 2020. Neural machine translation from jordanian dialect to modern standard arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*.

Sultan Alrowili and Vijay Shanker. 2022. Generative approach for gender-rewriting task with ArabicT5. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 491–495, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Laith Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). In *Computational Intelligence and Neuroscience.*

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

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. Investigating lexical replacements for arabic-english code-switched data augmentation.

Cong Duy Vu Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.

Saméh Kchaou, Rahma Boujelbane, Emna Fsih, and Lamia Hadrich-Belguith. 2022. Standardisation of dialect comments in social networks in view of sentiment analysis : Case of Tunisian dialect. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5436–5443, Marseille, France. European Language Resources Association.

Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich. 2023. Hybrid pipeline for building arabic tunisian dialect-standard arabic neural machine translation model from scratch. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(3).

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Arat5: Text-to-text transformers for arabic language generation. *arXiv preprint arXiv:2109.12068.*

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. Turjuman: A public toolkit for neural arabic machine translation. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT5)*, Marseille, France. European Language Resource Association.