AraT5-MSAizer: Translating Dialectal Arabic to MSA

Murhaf Fares
Independent Researcher
murhaf@proton.me

Abstract

This paper outlines the process of training the `AraT5-MSAizer` model, a transformer-based neural machine translation model aimed at translating five regional Arabic dialects into Modern Standard Arabic (MSA). Developed for Task 2 of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, the model attained a BLEU score of 21.79% on the held-out test set associated with the task.

Keywords: Arabic, Neural Machine Translation, T5

1. Introduction

Arabic—a Semitic language spoken by over 400M people—encompasses a range of languages and dialects that have varying degrees of mutual intelligibility (Bergman and Diab, 2022). Perhaps what is even more defining of the Arabic language is the state of *diglossia* where all regional and local Arabic dialects co-exist with a "very divergent, highly codified (often grammatically more complex) superposed variety" (Ferguson, 1959, p. 336)—which is the Modern Standard Arabic (MSA). MSA is often used in formal and legal contexts across Arab countries, while dialectal Arabic (DA) comprises a rich array of regional and local dialects, differing in phonology, morphology, syntax and semantics (Habash, 2022). These variations between Arabic dialects and MSA pose challenges for Arabic Natural Language Processing (NLP) systems, particularly because many of the existing datasets and corpora have been focused on MSA rather than the myriad of Arabic dialects, and the very fact that MSA is shared across the Arab world (Bender, 2019; Bergman and Diab, 2022).[1]

This paper presents a fine-tuned encoder-decoder model to translate dialectal Arabic into MSA. The model is the result of participating in Task 2 under the 6th Workshop on Open-Source Arabic Corpora and Processing Tools; the shared task is presented in more detail in Section 2. The model itself, along with the data used to train it, are described in Section 3. In Section 4 we report the results on the development and test datasets provided by the task organizers. We briefly refer to related work in Section 5 and reflect on findings and the way forward in Section 6.

2. Task Description

The Dialect to MSA Machine Translation Shared Task revolves around translating various Arabic dialects into Modern Standard Arabic, with the intention to bridge the gap between colloquial Arabic and formal written language. Participants were asked to develop models to accurately translate (or convert) dialectal Arabic into MSA. The task covered five regional dialects, namely: the Gulf, Egyptian, Levantine, Iraqi, and Maghrebi dialects. The development and test datasets provided in the task are modestly sized. The development set comprises $1,001$ sentence pairs—200 pairs per dialect—whereas the test set includes $1,888$ sentence pairs that are unevenly distributed over the dialects, as illustrated in Table 1.[2] Participants were allowed to utilize whichever resources available to train and/or fine-tune their systems. All submissions to the shared task were evaluated using two metrics, viz. BLEU (Papineni et al., 2002) and Comet DA (Rei et al., 2022).[3]

3. Model Description

We dubbed our model `AraT5-MSAizer`, and it is the result of fine-tuning the AraT5$_{v2}$ model by Nagoudi et al. (2022)—a pre-trained encoder-

[1] We suspect that there are political as well as religious factors contributing to the marginalization of dialectal Arabic, or even looking down at dialectal varieties as 'ill-formed' Arabic. Though not discussed any further here, it is imperative to examine the status of Arabic NLP resources in light of this, while acknowledging efforts like the OSACT 2024 Shared Task, among others.

[2] According to the Shared Task's website there was supposed to be 500 MSA-dialect pairs for each dialect, both for development and testing. "For each dialect, a set of 500 sentences written in both MSA and dialect will be provided for finetuning, and the testing will be done on a set of 500 blind sentences" https://osact-lrec.github.io.

[3] More details on the shared task and the results can be found on: https://codalab.lisn.upsaclay.fr/competitions/public_submissions/17118

Dialect	No. sentence pairs
Gulf	586
Levantine	568
Magharebi	343
Egyptian	314
Iraqi	77

Table 1: Dialect-wise breakdown of sentence pairs in the test dataset from the shared task.

decoder transformer model (Raffel et al., 2020).[4] We chose to fine-tune this specific model because it was pre-trained on Twitter data, among other datasets, which encompass dialectal Arabic (Nagoudi et al., 2022). In addition, as we describe in Section 5, the AraT5$_{v2}$ model has been used in other related shared tasks for dialect-to-MSA translation.[5] We approached the task as translation from dialect to MSA without distinguishing between the different dialects (even though those were provided in the development and test datasets).

In the following sub-sections, we present the training data used to fine-tune the model and the training configuration.

3.1. Training Data

To fine-tune our model, we used a blend of four distinct datasets; three of which comprised 'gold' parallel MSA-dialect sentence pairs. The fourth dataset, considered 'silver', was generated through back-translation from MSA to dialect, as detailed in Section 3.1.2.

3.1.1. Gold Data

The Multi-Arabic Dialects Application and Resources (MADAR). MADAR includes a parallel corpus of 25 Arabic city-level dialects in addition to MSA (Bouamor et al., 2018). As mentioned before, we train one model to translate from all dialects to MSA, and therefore we 'collapsed' all dialects and sub-dialects in MADAR to just DA, leading to a total of $88,200$ sentence pairs. We reserve an additional $9,800$ pairs for early evaluation and experimentation.[6] MADAR was also used in former related shared tasks such as the Nuanced Arabic Dialect Identification Shared Task organized by Abdul-Mageed et al. (2023).

The North Levantine Corpus. Krubiński et al. (2023) recently introduced a multi-parallel corpus focusing on the North Levantine dialect (aka the 'Shami' or Syrian dialect). The corpus is basically a subset of the OpenSubtitles2018 parallel corpora (Lison et al., 2018) where the Arabic sentences have been manually translated to the North Levantine Arabic dialect.[7] The corpus includes about $120,600$ Shami-MSA pairs; we used 90% of which for training.

The Parallel Arabic DIalect Corpus (PADIC). PADIC is a multi-dialect parallel corpus covering six Arabic (sub-)dialects of the Levantine and Maghrebi regional dialects (Meftouh et al., 2015, 2018). Like with MADAR, we do not distinguish between the different dialects for the purpose of training our model and, hence, end up with a dataset of $41,680$ dialect-MSA pairs.

3.1.2. Synthetic Data

One way to augment our training data is to exploit monolingual data (i.e. MSA-only datasets or corpora). Back-translation is an effective approach to 'create' more training data (Sennrich et al., 2016), where an MT system or model is trained in reverse; that is, the model is trained to translate target (MSA) to source (Arabic dialect). The resulting model can then translate target-side monolingual data back into the source language, creating a synthetic (or silver) parallel corpus for training a source-to-target model.

To generate the synthetic data, we first fine-tuned AraT5$_{v2}$ to translate from MSA into dialectal Arabic on the combination of the three aforementioned gold datasets.[8] We then used the resulting MSA-to-dialect model to translate a subset of the Arabic sentences in OPUS (Tiedemann, 2012; Zhang et al., 2020).[9] We filtered the sentences in OPUS to only include Arabic sentences that are longer than 5 characters and shorter than 450 characters.

Given the nature of the data in OPUS, some of the MSA-dialect pairs in the synthesized data included parentheses around foreign names in

[4] `AraT5v2-base-1024` is available on `https://huggingface.co/UBC-NLP/AraT5v2-base-1024`

[5] It is important to highlight that the model selection and training as well as the data creation process were also constrained by the limited resources available to the author as an independent researcher.

[6] We did not follow the original train-dev-test split in MADAR for selecting those sentences.

[7] The corpus includes pairings with several Indo-European languages but these are not relevant to the work presented here.

[8] We acknowledge that there isn't a singular entity called "dialectal Arabic". However, we posit that if the reverse-translation model is capable of producing any variation of dialectal Arabic, the reuslting synthetic corpus could prove beneficial.

[9] See: `https://huggingface.co/datasets/Helsinki-NLP/opus-100`

MSA, but not in the dialect translation; we post-processed the dataset to replace the opening and closing parentheses with the empty string in such cases.[10] The resulting synthetic parallel corpus consists of $965,020$ MSA-dialect pairs. As we will see in the following sub-section, not all of those pairs will be used for fine-tuning the final model.

One significant caveat of the MSA-to-dialect translation model is the dominance of the Levantine dialect, which is present in the three gold datasets used to train the model. Indeed the North Levantine Corpus is almost as large as PADIC and MADAR combined, and the last two already include Levantine sentences (cf. Table 2).

3.1.3. Training Dataset

The dataset used to train the model is the combination of the three gold datasets in addition to a further filtered version of the synthetic dataset. After the first round of experiments, we decided to filter out more sentence pairs from the synthetic dataset.

We used the MSA text length, again, to filter out all sentences that are shorter than 25 characters and longer than 300 characters. We opted to keep shorter sentences, as we observed the translation quality degrading as the sentence length increased. Lastly, we augmented the dataset with about $17,000$ randomly-selected sentences from MADAR where MSA is used as both the source and the target.[11] We included those instances to present the model with cases where no changes are required to 'transform' the source text into MSA.

The final combined dataset consists of $700,386$ dialect-MSA sentence pairs in its train split and $77,800$ pairs in the development split. Table 2 summarizes the size of the different datasets.

Dataset	**No. pairs**
MADAR	$88,200$
PADIC	$41,680$
North Levantine Corpus	$120,600$
Synthetic dataset - OPUS	$965,020$
Gold+synthetic†	$700,386$

Table 2: Number of dialect-MSA sentence pairs in the gold and synthetic datasets. † Gold+synthetic is the final combined and filtered dataset used to train the model.

3.2. Model Fine-tuning

We trained our models by fully fine-tuning AraT5$_{v2}$ for one epoch only using the Transformers library (Wolf et al., 2020). The maximum input length is set to 1024 (same as in the original pre-trained model) whereas the maximum generation length is set to 512. The learning rate and batch size were set to 2e-5 and 32, respectively.[12][13]

4. Results

To gauge the effect of fine-tuning on datasets of varying sizes and qualities, we fine-tuned three AraT5$_{v2}$ models:[14]

(1) `AraT5`$_{\texttt{MADAR}}$ trained on MADAR only

(2) `AraT5`$_{\texttt{Gold}}$ trained on the concatenation of the three gold datasets

(3) `AraT5`$_{\texttt{gold+synthetic}}$ trained on the gold and synthetic datasets

Table 3 shows result of evaluating the three models on the OSACT 2024 development split. From the table we clearly see that the model trained on both the gold and synthetic data outperforms the model trained on gold data only. This observation is consistent with the findings reported by Scherrer et al. (2023) regarding the effectiveness of back-translated data in enhancing the performance of their neural models. To understand how good (or bad) those models are we need a baseline 'model'. We simply used a leave-as-is baseline (Scherrer et al., 2023), where the dialect text is used as translation for MSA (i.e. copy the source to target) and attain 0.1445 in BELU score. With only MADAR data for fine-tuning, we end up with a lower performance than such a basic baseline approach.

As mentioned in Section 3.2, our models are trained for one epoch only, but we did evaluate `AraT5`$_{\texttt{gold+synthetic}}$ on the OSACT 2024 development set every $2,000$ steps. The result of this evaluation can be seen in Figure 1. Note that only greedy search was used with generation when evaluating on the development split. As can be seen from the figure, the model reaches its top performance (with 0.2325 in BLEU) after almost $15,000$ steps, but we don't restore the weights of the best performing model at the end training.

Even though we trained one model for all dialects, we can still examine the results per dialect, which are shown in Table 4.

[10] Parentheses are often used to enclose foreign names in Arabic (open) subtitles.

[11] On second thought, we think those examples could have been sampled from some other monolingual MSA resource.

[12] The training configuration as well as the training script can be found on https://github.com/Murhaf/AraT5-MSAizer

[13] The models were trained on one NVIDIA RTX A6000.

[14] All models were trained using the same configuration and (hyper)parameters outlined in Section 3.2

Model	BLEU
AraT5$_{\texttt{MADAR}}$	0.1140
AraT5$_{\texttt{Gold}}$	0.2038
AraT5$_{\texttt{gold+synthetic}}$†	0.2302
Baseline	0.1445

Table 3: BLEU score on the development split of the AraT5$_{v2}$ model fine-tuned on the MADAR dataset only, three gold datasets and the gold and synthetic datasets combined. † aka AraT5-MSAizer

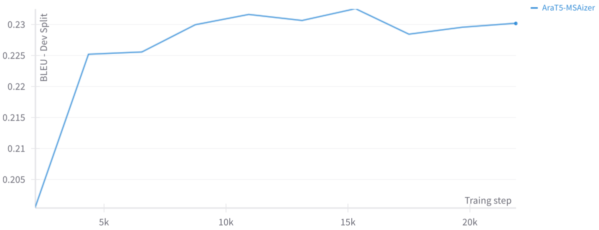

Figure 1: AraT5-MSAizer BLEU score on the OSACT 2024 development set every $2,000$ steps.

The results in Table 4 can be partly explained by the observation made by Bouamor et al. (2014) where they found that Egyptian had the highest lexical overlap with MSA while Tunisian had the least lexical overlap with MSA amongst all the dialects they studied.[15]

Lastly, Table 5 shows the official result of our fine-tuned model, AraT5-MSAizer, on the test split. We used beam search for the final translation submission (specifically, 6 beams) as beam search has proved to lead to better translation performance—at the cost of decoding speed though (Freitag and Al-Onaizan, 2017). Our BLEU score does seem reasonable compared to previously reported results on dialect-to-MSA translation (albeit on different evaluation datasets, cf. Section 5).

Dialect	BLEU
Egyptian	0.2708
Gulf	0.2373
Iraqi	0.2209
Levantine	0.2255
Magharebi	0.2087

Table 4: AraT5-MSAizer BLEU scores for the different dialects in the OSACT 2024 development set

Model	BLEU	Comet DA
AraT5-MSAizer	0.2179	0.0016

Table 5: Official evaluation results on the test split.

5. Related Work

There exists a substantial body of research on statistical and neural machine translation from DA to MSA, but in this section we only focus on Subtask 3 of the NADI-2023 Shared Task (Abdul-Mageed et al., 2023) as it is the most relevant to the OSACT 2024 Shared Task. Of the three participating teams, UniManc (Khered et al., 2023) and Helsinki-NLP (Scherrer et al., 2023) are the most similar to our approach. Both works—among other things—fine-tuned the AraT5$_{v2}$ model on existing parallel corpora for dialect-to-MSA translation. In addition, Scherrer et al. (2023) used a statistical machine translation model (SMT) to back-translate monolingual datasets into dialects which they then used as synthetic parallel corpora to train or fine-tune neural machine translation models.

UniManc—the winning team of task 3 in the NADI-2023 Shared Task—reached their best overall performance by fine-tuning the AraT5$_{v2}$ model on what they call "joint regional" configuration, where all dialect-to-MSA pairs were used to train the same model. We followed a similar approach in the work presented in this paper, but with the addition of synthetic data.

Helsinki-NLP achieved their best performance with SMT models. However, they also fine-tune the AraT5$_{v2}$ model on gold data (viz. MADAR) as well as synthetic back-translated data. Their findings are pretty much in line with ours in that fine-tuning on MADAR-only is barely enough and that back-translation can be effective in the context of fine-tuning pre-trained models.

6. Conclusion

In this paper we presented a machine translation model that builds on a pre-trained text-to-text language model to translate from five different Arabic dialects to MSA. We showed that we can utilize the already existing, though scare, parallel corpora to produce more training data from monolingual resources. We clearly demonstrated that such synthetic data (via back-translation) does indeed help boost the model's performance, in contrast to only relying on gold training data. Despite

[15] We checked the lexical overlap between MSA and the five dialects in the OSACT 2024 development set and found that Magharebi has indeed the least overlap. Note that our lexical overlap method is rather simple, we tokenized the source and target sentences in the dataset, computed the lexical overlap between each pair, and then averaged the lexical overlap per dialect.

the promising results showcased in this paper—which align with recent results in related tasks—we believe that back-translation is not exploited to its fullest yet. One pitfall we would like to avoid in future work is re-using the same 'genre' of text in the different datasets; this is especially the case for the North Levantive Corpus and the synthetic data we chose to back-translate. In addition, we believe one can try and test the idea of iterative back-translation (Hoang et al., 2018), but we suspect a better starting point for the reverse translation system is needed.

7. Bibliographical References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

A. Bergman and Mona Diab. 2022. Towards responsible natural language annotation for the varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland. 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, 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).

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Nizar Y Habash. 2022. *Introduction to Arabic natural language processing*. Springer Nature.

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

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.

Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospíšil, Petr Zemánek, and Pavel Pecina. 2023. Multi-parallel corpus of North Levantine Arabic. In *Proceedings of ArabicNLP 2023*, pages 411–417, Singapore (Hybrid). Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

Karima Meftouh, Salima Harrat, and Kamel Smaïli. 2018. PADIC: extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yves Scherrer, Aleksandra Miletić, and Olli Kuparinen. 2023. The Helsinki-NLP submissions at NADI 2023 shared task: Walking the baseline. In *Proceedings of ArabicNLP 2023*, pages 670–677, Singapore (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.