LLM-based MT Data Creation: Dialectal to MSA Translation Shared Task

AhmedElmogtaba Abdelaziz, Ashraf Elneima, Kareem Darwish
aiXplain Inc.,
San Jose, CA, USA
{ahmed.abdelaziz,ashraf.hatim,kareem.darwish}@aixplain.com

Abstract

This paper presents our approach to the Dialect to Modern Standard Arabic (MSA) Machine Translation (MT) shared task, conducted as part of the sixth Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT6). Our primary contribution is the development of a novel dataset derived from The Saudi Audio Dataset for Arabic (SADA), an Arabic audio corpus. By employing an automated method utilizing ChatGPT 3.5, we translated the dialectal Arabic texts to their MSA equivalents. This process not only yielded a unique and valuable dataset but also showcased an efficient method for leveraging large language models (LLMs) in dataset generation. Utilizing this dataset, alongside additional resources, we trained a machine translation model based on the Transformer architecture. Through systematic experimentation with model configurations, we achieved notable improvements in translation quality with BLEU scores advancing from a baseline of 25.5 to a peak of 31.5 in varied experimental setups. Our findings highlight the significance of LLM-assisted dataset creation methodologies and their impact on advancing machine translation systems, particularly for languages with considerable dialectal diversity like Arabic.

Keywords: Modern Standard Arabic, Dialectal Translation

1. Introduction

The field of neural machine translation (NMT) has seen remarkable progress in recent years. Yet, translating Arabic dialects to Modern Standard Arabic (MSA) presents unique challenges. These challenges stem from the vast linguistic diversity across Arabic dialects and the scarcity of dialect-specific corpora for training effective machine translation systems. Further, there is large lexical overlap Arabic dialects and MSA, and many dialects exhibit common syntactic properties. This paper details our approach to addressing these challenges, highlighting our participation in the Dialect to Modern Standard Arabic Machine Translation shared task at the sixth Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT6) (Elneima et al., 2024).

A pivotal aspect of our contribution lies in leveraging the Saudi Audio Dataset for Arabic (SADA) (Alharbi et al., 2024), an extensive Arabic audio corpus, as the foundation for generating a novel text-based dataset. We developed an automated method that employs ChatGPT 3.5 to translate dialectal Arabic text to corresponding MSA text. This process not only generates a substantial corpus of reasonable quality dialect-specific data but also demonstrates the potential of using Large Language Models (LLMs) for dataset creation in an automated and scalable manner.

Building on this foundation, we explored the efficacy of our newly created dataset, both independently and in conjunction with existing datasets, to train state-of-the-art transformer-based MT models (Vaswani et al., 2017). Our methodology encompasses a detailed examination of model configurations, focusing on optimizing attention heads and embedding dimensions to enhance translation accuracy and fluency. Our contributions are as follows:

- We show the efficacy of using LLMs (ChatGPT 3.5) for creating parallel dialectal-MSA data.
- We build a robust dialect to MSA MT system that combines both existing datasets and LLM generated data.
- We train a single transformer model that translates all dialects to MSA.

2. Related Work

In the evolving landscape of neural machine translation for Arabic dialects, research has predominantly been focused on bridging the linguistic divide between various regional dialects and Modern Standard Arabic. Despite the advancements, the challenge remains in developing comprehensive models that can accommodate the wide array of dialects spoken across the Arab world. In light of these challenges, our work draws inspiration from and seeks to build upon the foundation laid by previous studies, while introducing novel methodologies to enhance translation accuracy and efficiency.

OSACT 2024 Workshop, pages 112–116

The study by Al-Ibrahim and Duwairi (2020) on the Jordanian Arabic dialect and the investigation into Levantine dialects by Baniata et al. (2018) underscore the potential of deep learning techniques in dialect translation and highlight the limitations imposed by dataset size. Our approach similarly leverages deep learning while incorporating an innovative dataset expansion strategy using an existing dataset, namely SADA (Alharbi et al., 2024), combined with automated translation via LLMs, namely chatGPT 3.5, to overcome the corpus size limitation.

Moreover, the transductive transfer learning strategy employed by Yazar et al. (2023) for the Algerian Arabic dialect showcases the effectiveness of knowledge transfer between models. They utilize a pre-trained AraT5 transformer model as the backbone for their translation system. The introduction of TURJUMAN (Nagoudi et al., 2022) represents a significant leap forward, offering a versatile tool for translating multiple languages into MSA. Our work aligns with the spirit of TURJUMAN, emphasizing flexibility and the use of advanced deep learning models. However, we differentiate our approach by focusing on the automated generation of a good-quality dialect-specific dataset that can further refine the translation process.

Lastly, Kchaou et al. (2023) developed a hybrid model using JoeyNMT for the Tunisian dialect translation, achieving good results. We extend this concept by experimenting with various model configurations and training strategies leading to good results across multiple dialects.

3. Experimental Setup

3.1. Data

For training our dialectal to MSA MT systems, we used two different datasets, namely the NADI dataset (Abdul-Mageed et al., 2023) comprising a total of 124,000 segments across various Arabic dialects as detailed in Table 1, and a conversational dataset that we extracted from the SADA speech corpus (Alharbi et al., 2024) and automatically translated to MSA using chatGPT 3.5, which will henceforth refer to as SADA-DA. The dataset contains 1,027,153 segments of naturally occurring dialectal conversations with the breakdown per dialect shown in Table 2.

3.1.1. NADI Dataset

NADI dataset (Abdul-Mageed et al., 2023) is particularly notable for its diversity, encompassing a wide range of Arabic dialects from across the Arab world. The inclusion of the NADI dataset significantly enriched our training corpus, providing a broad spectrum of dialectal variations and linguistic nuances. It spans many Arabic dialects with their sub-dialects. Table 1 lists the dialects in the NADI dataset.

Dialect	Segments
Tunisian	14,000
Iraq	4,000
Libya	4,000
Morocco	14,000
Syria	4,000
Saudi Arabia	4,000
Egypt	4,000
Jordan	42,000
Palestinian	2,000
Qatar	12,000
Yemen	2,000
Algeria	2,000
Lebanon	12,000
Oman	2,000
Sudan	2,000

Table 1: Breakdown of NADI dataset

Dialect	Segments
Hijazi	690,784
Najdi	298,866
Egyptian	11,900
Levantine	7,542
Morrocan	5,540
Algerian	4,677
Janubi	3,603
Iraqi	2,683
Shamali	1,558

Table 2: Breakdown of SADA-DA dataset

3.1.2. SADA-DA

SADA is an Arabic audio dataset composed of roughly 650 hours that are transcribed, diarized, and annotated with gender, approximate age, and dialect (Alharbi et al., 2024). From the SADA dataset, we extracted the transcription of the audio segments that were marked as dialectal. One of the main advantages of the SADA dataset is that the segments are composed of naturally occurring dialectal conversations spanning many genres and topics. Table 2 shows the breakdown per dialect for the SADA-DA.

As can be seen from the dataset, Gulf dialects, namely Hijazi, Najdi, Janubi, and Shamali, are over represented. We prompted chatGPT 3.5 to produce their MSA equivalents. Here are some sample segments with their automatically generated MSA equivalents:

- Shamali:

- أبي الغدا الغدا بسرعة واللي يرحم والدينك
- أريد الغداء الآن بسرعة، وبمن يرحم والديك

- Najdi:
  - لا لا بس بس أغراضي خلها اسمع
  - لا لا، فقط أغراضي اتركها واستم

- Morrocan:
  - اه دبما كيقول هكذا
  - دائماً يقول هكذا.

An important note here is that since the validation and test sets for the shared task were also drawn from SADA, we made sure that none of our training sentences were in either set.

To guide the translation process and ensure consistency in the output, we crafted a specific prompt that directed ChatGPT 3.5 to translate texts into MSA, maintain the original text alongside its translation, and separate them using a designated symbol. The prompt used was as follows:

ترجم النصوص التالية للغة العربية الفصحى ،
اكتب كلا من النص الاصلي وترجمته بالعربية الفصحى
وافصل بينهما باستخدام هذا الرمز #

Translation: *Translate the following texts to MSA. Output the original text and its translation with a # as a separator between them.*

This approach allowed for the automated generation of good-quality parallel sentences, where the original dialectal Arabic text and its MSA translation were clearly delineated by the # symbol.

Our experiments illuminated the critical influence of prompt structure on the ChatGPT 3.5 output quality and the tendency to generate hallucinations or inaccurate content. It became evident that simplicity and clarity in prompt design were paramount. By formulating prompts that were succinct and to the point, we minimized the likelihood of hallucinations, thereby enhancing the reliability and accuracy of the translations produced by ChatGPT 3.5. This strategic approach to prompt crafting, focusing on brevity and directness, proved instrumental in facilitating more more accurate machine translations from dialectal Arabic to MSA.

We carried out a preprocessing step that looked into how the lengths of the original and translated texts varied. By spotting and excluding translations with major length discrepancies, we honed in on including only the most promising translations. Building upon this foundation, we proceeded with a manual review by assessing a randomly chosen sample of ChatGPT 3.5's translations, focusing on the translations' fluency. This step was important for identifying language subtleties that automated evaluations, such as BLEU scores, might miss.

3.2. Translation Model

We employed a transformer-based architecture (Vaswani et al., 2017) to address the challenge of translating dialectal Arabic to MSA. The Transformer model, renowned for its effectiveness in capturing complex dependencies in sequence-to-sequence tasks, consists of an encoder-decoder structure. Both the encoder and decoder comprise 6 layers, with each layer hosting 8 attention heads, facilitating the model's ability to focus on different parts of the input sequence simultaneously.

The embedding layers are post-processed with dropout, and the subsequent layers undergo dropout, addition, and normalization, enhancing the model's generalization capability. We employed a dropout rate of 0.1 and label smoothing of 0.1 to mitigate overfitting and improve the model's performance on unseen data.

We utilized tied embeddings, a technique that shares the weight matrix across the input and output embeddings and the decoder's pre-output layer, reducing the model's parameters and encouraging more semantic representations. We used an Adam optimizer with the hyperparameters: 0.9, 0.98, and 1e-09 and a gradient clipping norm of 5. The learning rate is set to 0.0003 with a warm-up of 16,000 steps followed by an inverse square root decay, facilitating a stable and effective convergence.

To facilitate the training of our translation model, we leveraged two state-of-the-art neural machine translation frameworks: Marian NMT (Junczys-Dowmunt et al., 2018) and JoeyNMT (Kreutzer et al., 2019). These frameworks are known for their efficiency, flexibility, and the high quality of the translation models they can produce.

4. Results

The effectiveness of our translation models was rigorously evaluated using the BLEU score (Papineni et al., 2002), a benchmark metric for assessing the quality of machine-translated text relative to a set of reference translations. Our evaluation strategy involved two sets of experiments to discern the impact of dataset composition on translation accuracy.

In the initial phase, we utilized SADA-DA exclusively. With the model configured with 4 attention heads and embedding dimensions of 256 for both the encoder and the decoder, we achieved a BLEU score of 25.5 on the validation set. This served

as a solid baseline, demonstrating the feasibility of our approach for the translation task.

Next, we increased the model's capacity, adjusting the number of attention heads to 8 and the embedding dimensions to 512 for both the encoder and the decoder. This resulted in a notable improvement in translation quality, with the BLEU score reaching 30.2 on the same validation set. This marked a significant performance boost, highlighting the advantages of expanding model capacity for this specific translation challenge.

In the second set of experiments, we combined SADA-DA and NADI datasets to train our models, aiming to reap the benefits of a richer, more diverse training corpus. Under the initial configuration (4 attention head – 256 embedding dimensions) with the combined datasets, the BLEU score improved to 27.3, while the enhanced configuration (8 attention head – 512 embedding dimensions) yielded a further improved BLEU score of 31.5. These results underscore the value of leveraging composite datasets to improve the model's understanding and translation of diverse Arabic dialects into Modern Standard Arabic.

The combination of the NADI and SADA-DA datasets to train our machine translation systems resulted in an approximate 1% enhancement in translation accuracy, as indicated by improved BLEU scores. This enhancement can be attributed to several factors related to the diversity and complementarity of the datasets:

- Increased Linguistic Diversity: The NADI dataset, with its text-based collection spanning various Arabic dialects, and the SADA-DA dataset, derived from conversational audio, collectively encompass a wide linguistic spectrum. This diversity introduces the model to a broader range of dialectal variations, idiomatic expressions, and syntactic structures, enabling it to learn more comprehensive translation patterns.

- Complementary Data Characteristics: The NADI dataset primarily focuses on textual data from digital platforms, which may include formal and semi-formal dialectal usage. In contrast, SADA-DA, being sourced from conversational speech, includes informal dialectal expressions and colloquialisms.

- Robustness to Variability: Training on a mix of text-based and speech-derived datasets exposes the MT system to variations in spelling, grammar, and usage across different contexts.

- Improved Generalization: The combination of datasets mitigates the risk of overfitting to the peculiarities of a single dataset.

- Data Augmentation Effect: The addition of the SADA-DA dataset effectively serves as a form of data augmentation, increasing the volume of training data. This augmentation is particularly beneficial for dialects that are underrepresented in text-based corpora.

Dataset	Heads	Embed	BLEU
SADA-DA*	4	256	25.5
SADA-DA*	8	512	30.2
SADA-DA+NADI*	4	256	27.3
SADA-DA+NADI†	8	512	31.5

Table 3: Experimental results on the validation set using SADA-DA alone and SADA-DA+NADI (*MarianMT, †JoeyNMT)

These experiments illustrate the positive impact of dataset diversity and model capacity on machine translation performance, particularly in the context of translating Arabic dialects to MSA. The advancements in BLEU scores from using the combined SADA-DA and NADI datasets reaffirm the importance of comprehensive and varied training data in developing effective translation models.

5. Conclusion

In this paper, we presented our participation in the OSACT6 shared task on the translation of Arabic dialects to MSA, leveraging state-of-the-art neural machine translation techniques. Our research introduced a novel approach to dataset creation and utilization, primarily focusing on the automated generation of a text corpus from the SADA dataset. This method highlights the efficacy of using LLMs for data generation and the potential of using audio corpora in enriching machine translation training sets.

Our experiments were methodically designed to assess the impact of dataset composition and model configuration on translation performance. The initial experiments, conducted using the SADA-DA dataset alone, set a solid baseline for our translation models. Subsequent experiments with enhanced model capacities further improved translation quality, as shown by the observed increases in BLEU scores. The integration of the SADA-DA dataset with the NADI dataset enabled our models to benefit from a richer and more linguistically diverse training sets. This combination led to notable improvements in BLEU scores, underscoring the value of diverse training corpora in the realm of machine translation. For future work, we plan to experiment with a greater variety of dialect-to-MSA parallel corpora and with n-shot prompting of LLMs.

6. 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.

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.

Sadeen Alharbi, Areeb Alowisheq, Zoltan Tuske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Alnajim Raneem Aloraini, Raghad, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. Sada: Saudi audio dataset for arabic. *ICCASP 2024*.

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

Ashraf Elneima, AhmedElmogtaba Abdelaziz, and Kareem Darwish. 2024. Osact6 dialect to msa translation shared task overview. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, LREC'2024*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

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).

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 1–11, Marseille, France. European Language Resources Association.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bılge Kağan Yazar, Durmuş Özkan Şahın, and Erdal Kiliç. 2023. Low-resource neural machine translation: A systematic literature review. *IEEE Access*, 11:131775–131813.