OSACT6 Dialect to MSA Translation Shared Task Overview

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

Abstract

This paper presents the Dialectal Arabic (DA) to Modern Standard Arabic (MSA) Machine Translation (MT) shared task in the sixth Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT6). The paper describes the creation of the validation and test data and the metrics used and provides a brief overview of the submissions to the shared task. In all, 29 teams signed up, and 6 teams made submissions to the competition's leaderboard, with five of them submitting papers to the OSACT6 conference. The teams used a variety of datasets and approaches to build their MT systems. The most successful submission involved using zero-shot and n-shot prompting of ChatGPT.

Keywords: Machine translation, Dialectal translation

1. Introduction

While **M**odern **S**tandard **A**rabic (MSA) serves as the standardized formal language across the Arab world, **D**ialectal **A**rabic (DA) encompasses various regional dialects with unique vocabulary and morphology. However, resources for processing DA are scarce, posing challenges for tasks like machine translation. To overcome this, researchers have explored methods such as using MSA as a bridge language for translation. By pivoting on MSA, the translation accuracy of highly dialectal Arabic text into other languages could be enhanced.

The dialect to MSA machine translation shared task offers an opportunity for researchers and practitioners to tackle the intricate challenge of translating various Arabic dialects into Modern Standard Arabic. With the rich linguistic diversity across Arabic-speaking regions, this task aims to advance machine translation capabilities and bridge the gap between colloquial spoken Arabic and the formal written language. Participants worked on developing and refining translation models that can accurately and fluently convert dialectal Arabic text into MSA, making it a crucial initiative for improving communication and comprehension in the Arabic-speaking world.

The shared task covers multiple dialects, namely: Gulf, Egyptian, Levantine, Iraqi, and Maghrebi. For each dialect, there is a set of 200 sentences written in both MSA and dialect will be provided for fine-tuning (validation set), and the testing was done on a blind set of 1,888 test sentences that cover all 5 dialects (test set). The participants were free to use whatever resources at their disposal to train and fine-tune their systems. In this paper we:

- Describe the dataset and metrics that were used
- Introduce the common approaches that the participants used in their submissions

The shared task was run on CodaLab, and the details of submissions, data formats, and leaderboard reside there[1].

2. Related Work

Several works focused on machine translation from dialectal Arabic to MSA. For instance, Guellil et al. (2017) proposed a neural system translating Algerian Arabic (Arabizi and Arabic script) to MSA, while Baniata et al. (2018) introduced a system for translating Levantine and Maghrebi dialects to MSA. The **N**uanced **A**rabic **D**ialect **I**dentification (NADI) (Abdul-Mageed et al., 2020, 2021, 2022, 2023) task series is dedicated to addressing challenges in general Arabic dialect processing. While the first two versions focused on dialect identification and sentiment, the 2023 edition emphasized machine translation from Arabic dialects to MSA, a critical yet relatively nascent NLP task. Subtasks 2 and 3 of NADI2023 focused on machine translation from four Arabic dialects (Egyptian, Emirati, Jordanian, and Palestinian) to MSA at the sentence level. The datasets for these subtasks, named MT-2023-DEV and MT-2023-TEST, were manually assembled. MT-2023-DEV consists of 400 sentences, with 100 representing each dialect, while MT-2023-TEST comprises a total of 2,000 sentences, with 500 from each dialect. For subtask 3 training, participants were given the freedom to use additional datasets, whereas subtask 2 was restricted to utilizing MADAR-4-MT only. The MADAR corpus contains parallel sentences representing the dialects of

[1] `https://codalab.lisn.upsaclay.fr/competitions/17118`

OSACT 2024 Workshop, pages 93–97

25 cities across the Arab world, with translations in English, French, and MSA (Bouamor et al., 2019a). Addressing the original dataset's lack of country-level labels, a mapping was executed to link the 25 cities to their respective countries, resulting in the creation of MADAR-18. Furthermore, MADAR-4-MT integrates dialectal-to-MSA data from four specific dialects (Egyptian, Emirati, Jordanian, and Palestinian) extracted from MADAR-18, tailored for training MT systems in subtask 2.

3. Data and Metrics

3.1. Data

To create the validation and test set, we extracted 2,000 random segments per dialect from the **S**audi **A**udio **D**ataset for **A**rabic (SADA), which is an Arabic audio dataset composed of roughly 650 hours that are transcribed and annotated with gender and dialect (Alharbi et al., 2024). For the Gulf dialect, SADA used finer-grained labels, namely Najdi, Hijazi, Gulf, Shamali, and Gulf. Thus, we combined all of them when picking the random samples. Similarly, we combined Algerian and Moroccan segments for the Maghrebi dialect. Given the randomly extracted samples, we followed a two-step process to translate them into MSA. First, we prompted chatGPT to translate the dialectal sentences to MSA using the following prompt:

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 standard Arabic. Write the original text followed by the standard Arabic and separate between with them with # symbol.*

In the second step, we enlisted the help of native speakers of the different dialects to review the translations to ascertain their correctness and to correct the translations as needed. The reviewers had the option of accepting the translation as is, editing and accepting, or skipping if: the source dialect was different, the source was MSA, or the source was not comprehensible or translatable. The reviewing was done using a version of Label Studio[2] on the aiXplain platform[3] with the interface shown in Figure 1. We asked the reviewers to review at least 500 segments. Table 1 shows the breakdown of the reviewed segments.

As can be seen, we surpassed 500 segments for all dialects except Iraqi. For all, we randomly picked 200 for validation and used the rest for testing. The validation set was provided with the ground truth translation, while the test set was provided without translation. Table 2 shows reviewed samples for the different dialects.

Dialect	Total	Valid	Test
Gulf	786	200	586
Levantine	768	200	568
Maghrebi	543	200	343
Egyptian	514	200	314
Iraqi	277	200	77

Table 1: The breakdown of the reviewed segments.

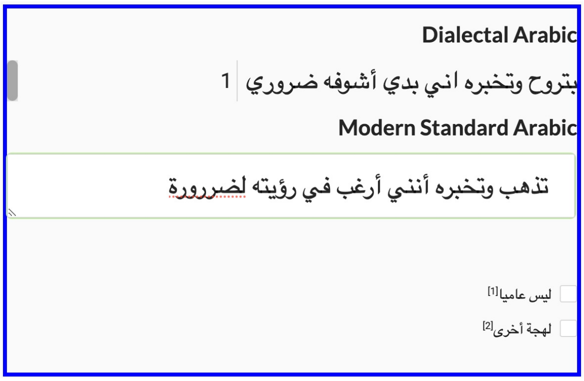

Figure 1: Reviewer interface

3.2. Metrics

For evaluation, we elected to use 2 different metrics that require ground-truth references, namely BLEU (Papineni et al., 2002) and Comet DA (Rei et al., 2022), which reportedly better correlates with human judgments compared to BLEU. While BLEU ranges between 0 and 1, with 1 being the highest possible score, Comet DA ranges between -1 and 1, with 1 being the highest score. BLEU was computed using the NLTK toolkit[4]. Since the computation of Comet DA is relatively computationally expensive, the computation was done on the aiXplain platform[5].

4. Submissions

Out of the 29 teams that signed up for the shared task, 6 teams made submissions. The teams used a variety of datasets and approaches to train their MT systems. Table 3 showcases the outcomes achieved by the participating teams.

MBZUAI (Atwany et al., 2024): The MBZUAI team used the MADAR dataset (Bouamor et al., 2019b) for training, which includes 95,600 dialectal

[2]https://labelstud.io/
[3]https://label.aixplain.com
[4]https://www.nltk.org/
[5]https://platform.aixplain.com

dialects	source	target
Gulf	عبد الله من جد يعني خاش	عبد الله دخل حقا
Egyptian	هتكون مين يعني العروسة؟	من ستكون إذا العروسة؟
Levantine	إي حركة لا تخليه لوحده	أي حركة لا تتركه وحده
Iraqi	هلا هلا والله بوخي وعليكم السلام عوافي عوافي يا وخي	مرحباً بك يا صديقي وعليكم السلام، أصابتك العافية
Maghrebi	ربي يهدينا ويرزقنا حسن الخاتمة يااارب	اللهم اهدنا وارزقنا خاتمة حسنة يا رب

Table 2: Random samples from the validation set

Group	BLEU	Comet DA
MBZUAI	29.6	0.028
aiXplain	25.2	-0.005
ASOS	22.3	0.004
MSAizer	21.8	0.002
nourrabih	10.1	-0.098
Sirius_Translators	9.6	-0.064

Table 3: Results for teams who submitted results and papers.

sentences with their corresponding MSA equivalents. The team experimented with a variety of models including the No Language Left Behind (NLLB) MT model from Meta, with and without finetuing, AraT5 with fine-tuning (Nagoudi et al., 2022), and chatGPT in zero-shot and 3-shot settings. Their team achieved the best results in the shared task using chatGPT prompting with 29.6 and 0.028 BLEU and Comet DA scores respectively. The $nourrabih$ team seems to have merged with the MBZUAI team.

aiXplain (Abdelaziz et al., 2024): The aiXplain team used two training datasets, namely the NADI dataset (124,000 sentences) (Derouich et al., 2023) and segments that were extracted from the SADA dataset and automatically translated to MSA using chatGPT 3.5 (1,027,153). For the MT model, they used two different neural MT toolkits, namely MarianMT (Junczys-Dowmunt et al., 2018) and Joey NMT (Kreutzer et al., 2019). Their best results were 25.2 and -0.005 for BLEU and Comet DA respectively on the test set.

ASOS (Nacar et al., 2024): The ASOS team employed data augmentation techniques utilizing GPT-3.5 and GPT-4 to increase the validation set size from 200 to 600 examples per dialect. They leveraged a dataset comprising 3000 samples (600 for each of the 5 dialects) for fine-tuning AraT5 v2. Their top-performing results on the test set were 22.3 for BLEU and 0.004 for Comet DA.

MSAizer (Fares, 2024): The MSAizer team fine-tuned the AraT5 model using four different datasets. Three of these datasets consisted of dialect to MSA pairs, namely: MADAR (95,600 sentences) (Bouamor et al., 2019b), NLC (120,600) (Krubiński et al., 2023), and PADIC (41, 680) (Meftouh et al., 2015). The fourth dataset was created by back-translating sentences from MSA, using a subset of OPUS data (965, 020) (Tiedemann, 2012). The final training dataset comprised 700,386 dialect-MSA sentence pairs. Their best results on the test set were 21.79 BLEU and 0.002 for Comet DA, respectively.

Sirius_Translators (Alahmari, 2024): This teams used 5 different datasets to train an MT model, namely MADAR (95,600 sentences) (Bouamor et al., 2019b), PADIC (32,060) (Meftouh et al., 2018), Dial2MSA (60,277) (Mubarak, 2018), Arabic STS (5,516) (Al Sulaiman et al., 2022), SADID (5,994) (Abid, 2020). For translation, the team fine-tuned multiple AraT5 models, namely AraT5 base, AraT5v2-base-1024, AraT5-MSA-Base, and AraT5-MSA-Small, with AraT5v2-base-1024 (Nagoudi et al., 2022) achieving the best results with 9.6 and -0.064 for BLEU and Comet DA respectively on the test set.

5. Conclusion

In this paper, we presented the dialectal Arabic to MSA translation shared task for OSACT6. The validation and test data for the shared task were prepared using a combination of LLM-based automatic translation and human verification and correction. In all, 29 teams signed up for the shared task, with 6 of them making submissions to the competition's leaderboard and 5 of them submitting system papers. Two main themes appeared in the submission, namely: using LLMs for data augmentation and creation, and finetuing either NMT models or LLMs (most notably AraT5) for translation. The best results were attained using LLMs, specifically chatGPT, using zero-shot and n-shot prompting.

6. References

AhmedElmogtaba Abdelaziz, Ashraf Elneima, and Kareem Darwish. 2024. Llm-based mt data creation: Dialectal to msa translation shared task. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, LREC'2024*.

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. *arXiv preprint arXiv:2310.16117*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. Nadi 2021: The second nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2103.08466*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2210.09582*.

Wael Abid. 2020. The sadid evaluation datasets for low-resource spoken language machine translation of arabic dialects. In *International Conference on Computational Linguistics*.

Mansour Al Sulaiman, Abdullah M Moussa, Sherif Abdou, Hebah Elgibreen, Mohammed Faisal, and Mohsen Rashwan. 2022. Semantic textual similarity for modern standard and dialectal arabic using transfer learning. *Plos one*, 17(8):e0272991.

Salwa Alahmari. 2024. Sirius_translators at osact6 2024 shared task: Fin-tuning arat5 models for translating arabic dialectal text to modern standard arabic. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, LREC'2024*.

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*.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. Osact 2024 task 2: Arabic dialect to msa translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, LREC'2024*.

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

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019a. The madar shared task on arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019b. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

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.

Murhaf Fares. 2024. Arat5-msaizer: Translating dialectal arabic to msa. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, LREC'2024*.

Imane Guellil, Faical Azouaou, and Mourad Abbas. 2017. Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter. In *The 31st pacific asia conference on language, information and computation paclic*, volume 31, page 2017.

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.

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.

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.

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.

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

Hamdy Mubarak. 2018. Dial2msa: A tweets corpus for converting dialectal arabic to modern standard arabic. *OSACT*, 3:49.

Omer Nacar, Abdullah Alharbi, Serry Sibaee, Samar Ahmed, Lahouari Ghouti, and Anis Koubaa. 2024. Asos at osact6 shared task: Investigation of data augmentation in arabic dialect-msa translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools, LREC'2024*.

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.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.