ASOS at OSACT6 Shared Task: Investigation of Data Augmentation in Arabic Dialect-MSA Translation

Omer Nacar, Serry Sibaee, Samar Ahmed,
Abdullah I. Alharbi, Lahouri Ghouti, Anis Koubaa
Robotics and Internet-of-Things Lab, Prince Sultan University
Faculty of Computing and Information Technology Rabigh, King Abdulaziz University
Riyadh 12435 Saudi Arabia, Jeddah 22254 Saudi Arabia
{onajar, ssibaee, lghouti, akoubaa}@psu.edu.sa
aamalharbe@kau.edu.sa, Samar.sass6@gmail.com

Abstract

The translation between Modern Standard Arabic (MSA) and the various Arabic dialects presents unique challenges due to the significant linguistic, cultural, and contextual variations across the regions where Arabic is spoken. This paper presents a system description of our participation in the OSACT 2024 Dialect to MSA Translation Shared Task. We explain our comprehensive approach, which combines data augmentation techniques using generative pre-trained transformer models (GPT-3.5 and GPT-4) with the fine-tuning of AraT5 V2, a model specifically designed for Arabic translation tasks. Our methodology has significantly expanded the training dataset, thus improving the model's performance across five major Arabic dialects, namely Gulf, Egyptian, Levantine, Iraqi, and Maghrebi. We have rigorously evaluated our approach, using the BLEU score, to ensure translation accuracy, fluency, and the preservation of meaning. Our results demonstrate the effectiveness of our refined data and models, achieving a BLEU score of 85.5% on the validation set and 22.6% on the blind test set, indicating a successful bridging of the gap between different dialects. However, it's important to note that while utilizing a larger dataset resulted in significantly higher evaluation BLEU scores, the performance on the blind test set was relatively lower. This observation underscores the importance of dataset size in model training, revealing potential limitations in generalization to unseen data due to variations in data distribution and domain mismatches.

Keywords: Machine Translation, Data Augmentation, BLEU Score, Arabic Dialects

1. Introduction

The Arabic language, characterized by its rich diversity of dialects, is the primary mode of communication for over 420 million individuals across the Middle East and North Africa. This linguistic landscape is distinguished by a phenomenon known as diglossia, wherein Modern Standard Arabic (MSA) coexists with various regional dialects (Qudah et al., 2017). As the formal variant, MSA is ubiquitously employed in official discourse, educational frameworks, and literary works across the Arab domain. Conversely, dialectal Arabic (DA) encompasses the myriad vernacular languages intrinsically linked to specific regions, encapsulating the essence of local identities and cultural intricacies.

The coexistence of MSA and DA within this linguistic ecosystem poses substantial challenges for machine translation. The pronounced variations in dialectal expressions, coupled with the scarcity of extensive parallel corpora essential for practical training, often culminate in suboptimal translation outputs when conventional models, predominantly trained on MSA, are utilized for DA content. This predicament underscores the critical need for translation methodologies tailored to accommodate the unique attributes of DA, enhancing accuracy and contextual relevance in this linguistically complex environment (Darwish et al., 2021).

In seeking to address the aforementioned challenges, we participated in the OSACT 2024 Dialect to MSA Translation Shared Task, which aims to evaluate the performance of translation models across five major Arabic dialects: Gulf, Egyptian, Levantine, Iraqi, and Maghrebi. The primary objective of this work is to test the efficacy of sequence-to-sequence translation models, particularly those using the Text-to-Text Transfer Transformer (T5) framework, in translating DA into MSA (Raffel et al., 2020).

Our participation in this shared task entailed pre-training specific models and carefully enhancing the training data for the above dialects. We used the dataset provided by the shared-task organizers for the training phase, after which our models were evaluated using the development and test sets (consisting of 500 unseen sentences for each dialect during the test phase). We conducted several experiments to evaluate the performance of the models comprehensively (Nagoudi et al., 2022); we also implemented different training settings to improve the results and accuracy of the translation between DA and MSA.

The subsequent sections are structured as follows: Section 2 reviews prior studies, Section 3 describes our proposed method, Section 4 details

OSACT 2024 Workshop, pages 104–111

our experimental result, and, finally, we conclude with a summarization of our main findings.

2. Related Works

Given the increasing need for effective communication across diverse cultures and global borders, it has become essential to establish systems that tackle the challenges of multiple dialects. However, ensuring precise and efficient translations has become increasingly complex. Therefore, our goal is to explore a variety of approaches to improve the effectiveness of translation systems, specifically for MSA and Arabic dialects.

Sghaier and Zrigui (2020) propose a machine translation system designed to translate Tunisian Dialect (TD) text into MSA through a rule-based methodology. The translation process comprises three key stages: morphological analysis and disambiguation, lexical and structural transfer, and morphological generation with spelling corrections, resulting in the output text in MSA. Sajjad et al. (2020) present a benchmarking effort for dialectal Arabic-English machine translation aimed at tackling the challenges encountered in low-resource machine translation, particularly concerning Arabic dialects. It introduces an evaluation suite designed as a standard for measuring the effectiveness of Arabic-English machine translation systems specialized in dialectal Arabic. By combining existing Arabic-English dialectal resources and generating new test sets, it provides a comprehensive evaluation framework, covering various dialect categories, genres, and levels of dialectal diversity. The study employs a transformer-based seq2seq model for this purpose.

Al-Ibrahim and Duwairi (2020) delves into the application of Neural Machine Translation (NMT) for translating the Jordanian dialect into MSA using deep learning techniques, specifically the RNN encoder-decoder model. The RNN encoder-decoder model proves to be effective in translating the Jordanian dialect into MSA, achieving a high accuracy rate for word-to-word translation and a lower accuracy rate for sentence translation. Additionally, Convolutional Neural Networks (CNN) are utilized to enhance translation accuracy. Moreover, the study (Moukafih et al., 2021) addresses the challenges of machine translation for six Arabic dialects: Tunisian, Algerian, Moroccan, Syrian, and Palestinian. It introduces the PADIC dataset, a parallel corpus of Arabic dialects and MSA. It presents a neural multi-task learning framework leveraging inter-dialectal relationships to achieve superior translation results.

Furthermore, Alzamzami and Saddik address challenges in translating Arabic dialects on social media by introducing a multi-dialectal Arabic-English dataset. It details the dataset construction process, emphasizing meticulous translator selection and cultural considerations. Additionally, it highlights deep learning-based translation models for four Arabic dialects, utilizing transfer learning and Transformer architecture for improved accuracy. The proposed dataset and models aim to address the limitations in current translation systems for Arabic dialects, particularly in informal social media contexts, spotlighting deep learning-powered translation models tailored for four distinct Arabic dialects: Gulf, Levantine (Shami), Iraqi, and Yemeni.

3. Methodology

In this section, we present a comprehensive approach for tackling the shared issue of translating different Arabic dialects into Modern Standard Arabic (MSA). Considering the wide range of linguistic variations among Arabic-speaking areas, our approach aims to improve translation models for precision and fluency while also bridging the gap between formal written Arabic and informal spoken Arabic. In order to do this, we have used a blend of sophisticated data augmentation methods and processes for fine-tuning that are especially suited to the distinctive qualities of the Arabic dialects—Gulf, Egyptian, Levantine, Iraqi, and Maghrebi. Our method improves the accuracy and consistency of dialect-to-MSA translation by utilizing the most recent developments in machine translation technology, such as the use of generative pre-trained transformer models.

3.1. Data Augmentation

A key component of our approach is data augmentation, which aims to significantly expand the variety and amount of training data available for optimizing our translation models (Shorten et al., 2021). The model's capacity to generalize across many dialects and linguistic subtleties, as well as the lack of sufficient training data, are major obstacles that must be overcome in order to successfully complete machine translation tasks.

3.1.1. Implementation of Data Augmentation

To implement our data augmentation strategy, we utilized a novel approach by incorporating the capabilities of generative pre-trained transformer models, specifically GPT-3.5 and GPT-4 models. These models were tasked with generating additional training examples from the original set of 200 sentences provided for each dialect. The augmentation process involved the following steps:

Source Sentence Preparation: For each source sentence in the provided dialectal Arabic datasets, we prepared a prompt designed to guide

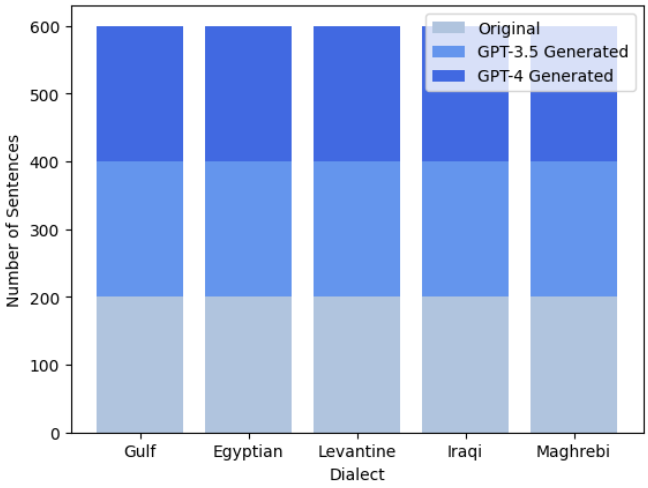

Figure 1: Dataset Size Before and After Augmentation by Dialect

the generative model towards producing a synonymous translation in MSA. The prompt explicitly instructed the model to ensure that the translation maintains the original sentence's meaning, adheres to Modern Standard Arabic grammar, and matches the original sentence in word count as closely as possible.

Model Interaction: We interacted with the GPT-3.5 and GPT-4 models through the OpenAI API [1], submitting each prepared prompt as input. The models were prefaced with a system message that outlined their role as language models trained for translating dialectal Arabic to MSA, emphasizing the need for accuracy, grammatical adherence, and word count maintenance.

Translation Generation: Upon receiving each prompt, the models generated translations that were then evaluated for quality and adherence to the specified criteria. This process allowed us to significantly expand our dataset with high-quality, model-generated translations, thereby enriching the training material available for fine-tuning our translation system. Figure 1. illustrates the dataset size before and after augmentation for each Arabic dialect.

3.1.2. Evaluation of Augmented Data

In assessing the quality of sentences generated by GPT models, traditional and advanced metrics provide insights into the linguistic and semantic fidelity of the output compared to target sentences. This evaluation highlights the challenges and solutions in quantifying the effectiveness of generative models in language tasks.

BLEU's Limitations in Sentence Evaluation, the Bilingual Evaluation Understudy (BLEU) metric, widely utilized in machine translation to measure the similarity of generated text to reference translations, showed significant limitations in our context of the evaluation step where BLEU evaluates the correspondence of n-grams between the generated and target texts, offering a score from 0 to 1. However, this method's reliance on exact matches often fails to capture the essence of semantic similarity and sentence structure, particularly in languages with rich morphology or when dealing with nuanced textual differences. A notable example from our dataset noted during evaluation illustrates this limitation in Figure 2,

As shown figure 2, Despite the generated sentence being semantically identical to the original target, except for the addition of a question mark, BLEU assigned a score of 0, demonstrating its inefficacy in capturing semantic equivalence and punctuation nuances.

Advantages of METEOR in Overcoming BLEU's Shortcomings, on the other hand and due to BLEU score sensitivity, the metric for evaluation of GPT models predictions are underscored with Explicit Ordering (METEOR) which offers a more nuanced evaluation by accounting for synonymy and stemming, in addition to exact matches. METEOR's alignment-based approach, which allows for a flexible matching of words and phrases, provides a more comprehensive assessment of similarity between the generated text and the target. Employing METEOR in our evaluation of GPT generated sentences yielded scores that more accurately reflected the semantic and syntactic correspondence between the target and GPT4 generated sentences as shown in Figure 3.

The average METEOR score across GPT4 and GPT3.5 augmented dataset are 73.22% and 67.48% respectively, indicating a strong alignment with the original ground truth MSA target sentences

[1] https://openai.com/blog/openai-api

Original Target	Generated Target	BLEU Score
كيف تتعلم How do you learn	كيف تتعلم؟ How do you learn?	0.00

Figure 2: BLEU Score Evaluation Demonstrating Sensitivity to Punctuation.

Original Target	Generated Target	METEOR Score
كيف تتعلم How do you learn	كيف تتعلم؟ How do you learn?	63.92
هل تعتقدين أننا سنصبح مثلهم في يوم من الأيام؟ Do you think we will be like them one day?	تعتقدين أننا سنكون مثلهم في يوم من الأيام؟ Do you think we will become like them one day?	81.17
نعم، لا يصدق Yes, unbelievable	أه لا يصدق Oh, unbelievable	72.31
أمي قالت لي لا بأس My mother told me it was okay	أمي قالت لي عفواً My mother said excuse me	45.36

Figure 3: METEOR Scores for Evalation of GPT-Enhanced Data

of similarity and the ability of METEOR to capture nuanced linguistic features.

Qualitative Evaluation with GPT-4 , In addition to quantitative metrics, we employed GPT-4 for a qualitative evaluation of sentence similarity. Using a custom prompt, sentences were assessed on a scale from 1 to 5, with 5 indicating identical semantic content. This approach allowed us to incorporate contextual understanding and nuanced judgment beyond the capability of automated metrics. Selected examples from our evaluation of GPT4 generated sentences are shown in Figure 4.

The average similarity score across evaluated pairs for GPT4 and Gpt3.5 are 4.59 and 4.43 respectively, demonstrating the efficacy of GPT-4 in understanding and evaluating semantic nuances.

Through evaluating GPT-3.5 and GPT-4 generated sentences, we harnessed their high-quality outputs for data augmentation, significantly boosting the AraT5 V2 machine translation performance from dialect to MSA. This approach effectively enriched our training dataset, showcasing the value of leveraging advanced language models in enhancing machine translation tasks.

3.2. Fine-Tuning AraT5-V2 for Enhanced Performance

Following the strategic data augmentation outlined in the previous section, we transition to the fine-tuning of AraT5 V2, a process central to our methodology aimed at enhancing Arabic dialect to MSA translation. AraT5 V2, the successor to the foundational AraT5 model, embodies a series of substantial upgrades that elevate its capabilities in Arabic language translation tasks significantly.

AraT5 (Nagoudi et al., 2022) is based on the same architectural foundation as the original T5 model, but trained solely on Arabic data comprising both MSA and dialectal Arabic (tweets) resulting in 29 Billion token with more than 248 GigaBytes of dataset. The most recent version of AraT5, AraT5 V2 was utilized in this work. A key improvement in AraT5 V2 lies in its training across a broader and more diverse Arabic data corpus. AraT5 V2 enhances the model's sequence length capability from 512 to 1024 tokens, doubling its capacity for handling longer text passages, ensuring context preservation and resulting in more accurate and coherent translations.

In order to evaluate the effectiveness of this model in our paper, we compare AraT5 V2 against different sequence to sequence machine translation models, including the ARaT5-base (Nagoudi et al., 2022), mT5 (Xue et al., 2020) models, to showcase the efficacy of AraT5 V2 in translating dialectal Arabic to Modern Standard Arabic (MSA). This benchmarking underscores why AraT5 V2 was the optimal choice for our study, highlighting its superior performance over the augmented dataset and specific advantages in addressing the complexities of dialect-to-MSA translation tasks. Table 1 illustrates the comparative analysis showing the validation loss and BLEU under the same training

Original Target	Generated Target	Similarity Score
كيف ذلك؟ How is that?	ماذا يعني ذلك؟ How is that?	3.5
كيف صحتك؟ How is your health?	كيف الصحة How is health	4.75
نعم، أعرفها Yes, I know her	أجل، أعرفها Yes, I know her	5

Figure 4: GPT4 Sentence Similarity Evaluation, Highlighting Semantic Alignment.

Model name	Validation loss	Validation BLEU
AraT5 V2	2.523	**0.255**
mt5	1.932	0.174
AraT5 Base	3.441	0.113

Table 1: Validation Loss and BLEU Scores for AraT5 V2, mt5, and AraT5 Base

configuration of all models.

As shown in Table 1, three models were evaluated based on their validation loss and BLEU scores: AraT5 V2, mT5, and AraT5 Base. AraT5 V2 demonstrated a compelling balance of performance metrics, recording a validation loss of 2.523 and a BLEU score of 0.255. Although mT5 presented a lower validation loss at 1.932, its BLEU score of 0.174 was notably inferior to that of AraT5 V2, indicating less effective translation quality. AraT5 Base, Although a key model, AraT5 Base had the highest validation loss of 3.441 and the lowest BLEU score of 0.113, putting it behind the others. These results clearly support chosen AraT5 V2 for our experiment, not only due to its superior BLEU score, which maintains a satisfactory balance between loss and translation quality, proving its possibility in handling the translation of dialect-to-MSA.

3.3. Training Configuration

The fine-tuning of AraT5 V2 is done by using two NVIDIA A100 GPUs for efficient large-scale machine learning tasks. The model was based on the $UBC-NLP/AraT5v2-base-1024$ model from hugging face, which is specifically designed for Arabic language tasks. The training used 128 tokens for source and target texts, a per-device batch size of 16, and 22 epochs to adapt the model without overfitting. The learning rate was 5e-5, using the AdamW optimizer, reflecting best practices in transformer-based models for NLP tasks.

Training was conducted on a dataset comprising 2,666 examples, with a validation set of 297 examples, ensuring the model's performance was evaluated. The dataset was split from a larger corpus, incorporating a diverse range of Arabic dialects and ensuring a comprehensive representation of linguistic nuances.

The model's performance was primarily evaluated using the $BLEU$ score, a widely recognized metric in machine translation that assesses the correspondence between the model's output and the target translations. This metric, coupled with our dataset, provided a robust framework for assessing translation quality and model effectiveness.

The AraT5 V2 model have been tested a thorough evaluation on a test set of 500 blind sentences after its training and fine-tuning phases, as part of the OSACT 2024 shared task. These sentences, representing a broad spectrum of Arabic dialects, provided a robust benchmark for testing the model's translation abilities. The evaluation, conducted blindly by the shared task organizers, primarily utilized the BLEU score to assess translation quality, focusing on accuracy, fluency, and meaning preservation.

The AraT5 V2 model's performance was comprehensively assessed through supplementary experiments, including augmenting the training dataset with dialectical variations like MADAR and evaluating its performance on synthetically generated datasets generated by GPT4 without fine-tuning, contributing to a comprehensive assessment of its efficacy across various real-world translation scenarios.

4. Evaluation and Results

Our experiments spanned a range of scenarios, each designed to evaluate different factors of model behavior and performance. We explore the impact of dataset size, data augmentation techniques, and fine-tuning strategies on model performance, lever-

Experiment ID	Experiment Type	Training Dataset	Dataset Size	Number of Steps	Val loss	Val BLEU
1	Dev Only FT	Dev Only	1k	5k	3.567	0.234
2	Dev Only FT	Dev Only	1k	10k	4.526	0.254
3	Madar + Dev FT	Madar and Dev	80k	85k	0.194	**0.855**
4	GPT4 Generated	Test Dataset	1k	-	-	-
5	Augmented Data with GPT4 + Dev FT	Dev + GPT generated	2k	4.5k	2.228	0.248
6	Augmented Data with GPT4 + Dev FT	Dev + GPT generated	2k	6k	2.523	0.255
7	Augmented Data with GPT3.5 + GPT4 + Dev FT	Dev + GPT generated	3k	2k	1.658	0.241
8	Augmented Data with GPT3.5 + GPT4 + Dev FT	Dev + GPT generated	3k	4k	1.732	0.237

Table 2: Summary of Experiments Results - Evaluation Metrics

Experiment ID	Experiment Type	Training Dataset	Dataset Size	Number of Steps	Test BLEU
1	Dev Only FT	Dev Only	1k	5k	0.215
2	Dev Only FT	Dev Only	1k	10k	0.215
3	Madar + Dev FT	Madar and Dev	80k	85k	0.172
4	GPT4 Generated	Test Dataset	1k	-	0.171
5	Augmented Data with GPT4 + Dev FT	Dev + GPT generated	2k	4.5k	0.222
6	Augmented Data with GPT4 + Dev FT	Dev + GPT generated	2k	6k	**0.226**
7	Augmented Data with GPT3.5 + GPT4 + Dev FT	Dev + GPT generated	3k	2k	0.205
8	Augmented Data with GPT3.5 + GPT4 + Dev FT	Dev + GPT generated	3k	4k	0.208

Table 3: Summary of Experiments Results - Test Metrics

aging both synthetic and real-world data sources. Additionally, we provide an error analysis framework to further understand the predictions and their limitations. All these experiment results are chosen based on the best epoch results of both validation loss and BLUE and they are fully summarized in Table 2 for validation set and Table 3 for blind test set.

Augmentation Method Effectiveness: Experiments 1 and 2 demonstrate pre-augmentation outcomes, achieving a 21.5% score post-training over 5k and 10k steps, respectively. With augmenting the training data with GPT4-generated samples (Experiments 5 and 6) demonstrated notable improvements in both evaluation and test BLEU scores achieving 22.6% as best score among others and compared to the baseline. This suggests that augmenting the dataset with diverse synthetic data can effectively enhance the model's performance, potentially by exposing it to a wider range of linguistic variations and nuances.

Impact of Dataset Size: The study, Experiment 3, used the larger Madar dataset Bouamor et al. (2018) and development dataset to achieve an impressive evaluation BLEU score of 85.5%. However, this performance did not extend to unseen test sets, where the score dropped to around 17.1%. The high score was observed when the 80K dataset was divided into training and validation sets, suggesting overfitting or overlap. The study highlights the importance of dataset composition and partitioning in model training, as larger datasets may not predict effectiveness on unseen data due to potential domain mismatches or differences in data distribution.

Untuned GPT-4 Translation Performance, experiment 4, which utilizes predictions directly generated by the GPT-4 model without any fine-tuning has acheived a BLEU score of 17.1%, surprisingly yields results comparable to those achieved with fine-tuned models. This observation suggests GPT-4's inherent capability to understand and translate Arabic dialects, underscoring its potential even in the absence of task-specific optimization.

Balancing Data Augmentation and Fine-tuning Experiments 7 and 8, which combined data from GPT3.5 and GPT4 for augmentation, yielded mixed results. While the evaluation BLEU improved compared to the baseline, the test BLEU scores did not show significant improvement. This suggests that a careful balance between data augmentation techniques and fine-tuning strategies is necessary to achieve optimal performance across various datasets and evaluation metrics.

GPT-4-Driven Error Analysis and Feedback, in our evaluation framework, we implemented a concise error analysis using four metrics—lexical, syntactic, semantic, and orthographic—to assess the translation quality from Arabic dialects to MSA. By utilizing GPT-4, we analyzed generated translations for adherence to the original sentences' meaning and structure, facilitating a targeted assessment of model performance across diverse dialects. This methodology enabled us to isolate areas of excellence and deficiency within each model, providing specific feedback on critical sentences representative of each dialect.

This strategic approach underscores the pivotal role of nuanced linguistic analysis in refining translation models, setting a foundation for subsequent enhancements. Figure 5 shows some samples of the performance of our translation models on selected sentences for Experiments IDs of 3, 6 and 8 which show better results among others.

As shown in Figure 5, the error analysis of Arabic dialect experiments reveals that GPT-4 models consistently maintain high fidelity to the original sentences' semantic content, syntactic structure,

Source			Predictions			Error Anaylsis Type				Feedback
ID	Dialect	Source	Experiment 6	Experiment 3	Experiment 8	Lexical Errors	Syntactic Errors	Semantic Errors	Orthographic Errors	-
418452	Egyptian	"وبعدين معناه ايه السؤال ده ده حتى اللي قتل لو جي حد يسأله" What does this question" mean, even if someone "?who killed comes to ask	"ثم ماذا يعني هذا السؤال، حتى لو كان القاتل إذا جاء أحد يسأله؟" br> "Then what does> this question mean, even if the killer comes to "?ask	"إذن ، ماذا عن سؤال آخر لم يطرحه أحد br> "So, what about> من قبل ؟" another question that no one "?has ever asked before	"ثم ما هو معنى هذا السؤال، هذا حتى لو جاء شخص يسأله؟" br> "Then what is the> meaning of this question, even if someone comes to ask "?it	Exp 6 and 8 accurately preserve meaning; Madar's translation in Exp 3 significantly deviates, introducing an unrelated scenario.	Exp 6 and 8 match source syntax; Madar in Exp 3 correct but deviates with new meaning.	Exp 6 and 8 capture original semantics; Madar in Exp 3 introduces unrelated concept, missing the source's essence.	No spelling errors; All Exp use punctuation correctly, despite Madar's semantic shift.	Exp 6 and 8 align closely with source; Madar in Exp 3 significantly diverges, possibly misinterpreting context.
418437	Egyptian	"اه ناقصهم حتة إلتزام شوية" "Ah, they lack a bit of commitment"	"نعم، ينقصهم بعض الالتزام قليلا" "Yes, they lack some commitment a little"	"نعم ولكن ببعض التخفيف في المصاريف br> "Yes, but with some> ". ".reduction in expenses	"نعم، ينقصهم شيء بسيط." "Yes, they lack something small."	Exp 5 shows accurate translation, Madar in Exp 3 shows incorrect context and Exp 8 shows somewhat curate, lacks specificity.	Structure matches source in Exp 3, note in Ex 3: Structure okay, but unrelated content and in Exp 8 ,simplified but correct structure	Exp 5 captures original meaning but in Exp 3 it misinterprets source, unrelated. and in Exp 8 shows some misses nuance such as "التزام".	Correct usage for all experiments	Exp 5 closely aligns with the source's intent and wording. Exp 3 introduces unrelated concepts, Exp 8 provides a somewhat accurate translation but lacks in specificity.
419298	Iraqi	"أكل أمي آمنة ذكرني بأمي" "My mother Amina's cooking reminds me of my mother"	"أكل أمي آمنة ذكرني بأمي" "My mother Amina's cooking reminds me of my mother"	"طعام أمي آمن ، ذكرني إلى أمي" . "My mother Amina's food, remind me of my mother."	"أكل أمي آمنة ذكرني بأمي" "My mother Amina's cooking reminds me of my mother"	Exp 6 and 8 Exact match. Madar in Exp 3: Alters "آمنة" to "آمن," changing meaning.	All models maintain original structure effectively.	Exp 6 and 8 Preserve meaning. Madar in Exp 3: Shifts "آمنة" to "safe," losing specific reference.	No errors detected across models; punctuation and spelling are correct.	Exp 6 and 8 accurately replicate the source. Madar's in Exp3 semantic shift from "آمنة" to "safe" alters the intended personal reference, diverging from the original meaning.
420067	Levantine	"إحنا كلياتنا مضغوطين أرجوكم" "We're all under pressure, please"	"نحن طلابنا مضغوطين، أرجوكم" "Our students are under pressure, please"	"كل منا مشغول على حدة ، من فضلك ". br> "Each of us is busy on> ".their own, please	"نحن طلابنا مضغوطون أرجوكم" "Our students are pressured, please"	Exp 6 and 8 : Accurate. Madar in Exp 3 Changes context to individual busyness.	Exp 6 and 8 : Structure matches source. Madar in Exp 3 Alters structure to emphasize individuals.	Exp 6 and 8 : Maintain collective stress context. Madar in Exp 3 Misinterprets collective stress as individual tasks.	No orthographic errors in any model; all maintain correct spelling and punctuation.	Exp 6 and 8 accurately translate the collective stress, matching the source's intent. Madar in Exp 3 shifts focus to individual busyness, diverging from the collective emphasis of the original sentence.
426600	Magharebi	"ايه عاقلة يا بعدي هههههههههه " br> "You're polite, my> "dear hahaha	"أيه عاقلة يا بعدي هههههههههه " br> "You're polite, my> "dear haha	"ماذا عن التي تليها ؟ هاها ، هاها ". br> "What about the one> ".that follows? Haha, haha	"نعم، عاقلة يا بعدي هههههههههه br> "Yes, polite, my> " "dear hahaha	Exp 6 and 8 Lexically faithful to source. Madar in Exp 3: Deviates significantly, altering context.	Exp 6 and 8 Preserve original structure. Madar in Exp 3 Changes sentence structure entirely.	Exp 6 and 8 Captures the playful tone. Madar in Exp 3 Loses original's affectionate tone, introduces unrelated query.	All Exp: Minor variation in laughter representation. Correct usage.	Exp 6 and 8 closely align with the source, maintaining its playful and affectionate tone. Madar in Exp 3 introduces an unrelated concept, diverging from the original sentiment.
50550	Gulf	"بس أشوفكم بعدين سلام" "Just see you later, goodbye"	"فقط أراكم ثم سلام." "Just see you then, goodbye."	"سأعود إليك عندما تكون أقل انشغالا" . "I will come back to you when you are less busy."	"ولكن أراكم لاحقا سلام" "But I'll see you later, peace"	Exp 6 and 8 Minor adjustments, remain faithful. Madar in Exp 3: Adds context not present in the original.	All models present a simplified but correct structure.	Exp 6 and 8 Retain the casual farewell intent. Madar in Exp 3 Shifts focus, implying a return during less busy times.	Correct spelling and punctuation across all models.	Exp 6 and 8 accurately capture the casual, farewell nature of the original. Madar, in Exp 3 however, introduces a nuance of returning during a less busy time, diverging from the direct farewell intention.

Figure 5: Sample Error Analysis Across Models, Highlighting Linguistic Accuracy and Model Performance.

and lexical choice, demonstrating their ability to translate Arabic dialects to MSA with minimal errors. However, Experiment 3 (Madar) often diverges from the source, indicating a potential gap in capturing the original's intent. The study emphasizes the importance of model selection in achieving high-quality translations of Arabic dialects and suggests targeted improvements for models struggling with semantic fidelity.

5. Conclusion

In our study for the OSACT 2024 Shared Task on translating Arabic dialects to MSA, we leveraged AraT5 V2 and data augmentation techniques with GPT-3.5 and GPT-4, achieving our best BLEU score of 22.6% with AraT5 V2. This underscores AraT5 V2's effectiveness in capturing the linguistic intricacies of Arabic dialects. Our error analysis further illuminated the strengths of GPT-4 in enhancing translation accuracy across lexical, syntactic, semantic, and orthographic dimensions. These results not only demonstrate the power of AraT5 V2 in handling Arabic translation tasks but also the importance of nuanced error analysis in refining model performance. Moving forward, we aim to integrate emerging technologies to push the boundaries of machine translation for Arabic dialects and MSA.

6. Acknowledgements

The authors thank Prince Sultan University for their support

7. Bibliographical References

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.

Fatimah Alzamzami and Abdulmotaleb El Saddik. 2023. Osn-mdad: Machine translation dataset for arabic multi-dialectal conversations on online social media. *arXiv preprint arXiv:2309.12137*.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Youness Moukafih, Nada Sbihi, Mounir Ghogho, and Kamel Smaïli. 2021. Improving machine translation of arabic dialects through multi-task learning. In *International Conference of the Italian Association for Artificial Intelligence*, pages 580–590. Springer.

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.

Mahmoud Ali Qudah et al. 2017. A sociolinguistic study: Diglossia in social media. In *Conference Proceedings. Innovation in Language Learning 2017*.

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.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

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

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

8. Language Resource References

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)*.