World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:11, 2024

Fine-Tuned Transformers for Translating Multi-Dialect Texts to Modern Standard Arabic

Tahar Alimi, Rahma Boujebane, Wiem Derouich, Lamia Hadrich Belguith

***Abstract*—**Machine translation task of low-resourced languages such as Arabic is a challenging task. Despite the appearance of sophisticated models based on the latest deep learning techniques, namely the transfer learning and transformers, all models prove incapable of carrying out an acceptable translation, which includes Arabic Dialects (AD), because they do not have official status. In this paper, we present a machine translation model designed to translate Arabic multidialectal content into Modern Standard Arabic (MSA), leveraging both new and existing parallel resources. The latter achieved the best results for both Levantine and Maghrebi dialects with a BLEU score of 64.99.

***Keywords*—**Arabic translation, dialect translation, fine-tune, MSA Translation, transformer, translation.

I. Introduction

ALTHOUGH it is known by its diglossia and ambiguity, the Arabic language today is making considerable progress in several areas of NLP, and especially machine translation. Generally, Arabic is divided into two large families: MSA which is the formal language used in education, governmental services and official documents and written literature, and AD which are the languages naturally spoken by Arabs. Over time, and as a mother tongue, AD become the main language used in all communications namely media, social networks and websites which leads to the need for automatic translation systems to facilitate the understanding of each dialect message.

Chronologically, researchers began by setting up various systems of translation from MSA as a formal language into other languages such as English and French. Subsequently, new approaches that attempt to translate some varieties of AD into MSA, English, etc. have appeared. But we have not had models capable of jointly translating AD.

AD differ from one region to another even within the same country, for example, in Tunisia, the southern dialect differs widely from the coastal dialect. The varieties and complexity of Arabic dialects made it very hard to build multidialectal translation models, hence the appearance of several researches that where focused on the translation of a simple Arabic dialect such as [14] and [1]. To overcome this problem, we must focus on the fact that such a geographical area has several terms in common, hence the possibility of treating AD collectively by groups: Gulf dialects, Levantine dialects, Maghreb dialects, and Egyptian dialect.

Although the MSA is believed to be the mother language of AD, the latter present several differences at the semantic, phonological and morpho-syntactic levels. Therefore, most machine translation systems designed for MSA cannot be easily adapted to AD and give low scores. This linguistic variation challenges the creation of automatic translation tools for AD to achieve better performance.

The goal of the current work is to develop a high-level multidialectal machine translation system into MSA. For this, we built parallel resources that we have mixed with existing parallel corpora. Then, we developed three multidialectal MT models; the first translates the Levantine dialects into MSA, the second translates the Maghreb dialects into MSA and the last is the fusion of the above two models which translates collectively all dialects (Levantine and Maghreb) into MSA. In this paper, the proposed models involve the fine-tuning of T5X [23] and AraT5v2-base-1024 [19], which is the last version of AraT5 transformer. The models were evaluated in two ways: on dialects included in the training dataset, and on a zero-shot translation between language pairs never seen explicitly during the training phase. All experiments showed significant results and revealed the effectiveness of the chosen approach.

The rest of the paper is organized as follows: in Section II, we review related works in the literature evaluating transfer transformers in the machine translation field of AD. The external parallel data exploited to build and fine-tune the transfer transformers are described in Section III. Section IV presents the proposed multidialectal MT models and our different settings. We provide official results and evaluation on Section V. In Section VI, we summarize our proposed work. Finally, we conclude and we discuss future directions in Section VII.

II. Related Works

Machine translation is a challenging task for Arabic language, especially for AD, due to its complexity and the big differences compared to MSA. Based on the source and target language, we can classify the MT approaches into three main classes. The first approach focuses on translating AD into MSA; the second approach aims to translate from AD into other formal foreign languages such as English, while the last approach focuses on the interdialectal translation, whether directly or via an intermediary language.

Attempts to translate AD started from classical system, namely rule-based approaches [16], [17], [25] and empirical approaches [8], [10], [18], to recent new systems based on transfer learning and fine-tuning of transfer transformers.

T. A., R. B., W. D., and L. B. are with the ANLP Research group - MIRACL Lab., University of Sfax, Tunisia (e-mail: alimitahar2020@ gmail.com, rahma.boujelbane@fsegs.usf.tn, wiemderwich123@gmail.com, lamia.belguith@fsegs.usf.tn).

Despite the lack of a perfect model, approaches based on transformers are considered the most efficient and have significantly advanced the machine translation task.

The weak point is that initially, the automatic translation models were based on monolingual corpora. To overcome this weakness, researchers built parallel corpora which represent a text in two languages, source and target [26], [10], [14]. The use of parallel resources optimized the results of the MT systems a little, but we still need a perfect automatic translation system. Over time, transformers, which are advanced models pretrained on a large huge of data, appeared to overcome all insufficiencies. The last transformers built for Arabic translation were pretrained on parallel corpora that cover both MSA and its dialect varieties such as AraT5v2-base-1024, which is an extension and an optimization of the Arabic text-to-text transformer AraT5 [19]. Some researches tried to translate from a single dialect into MSA or English language. Reference [14] achieved a BLEU score of 60.00 in translating the Tunisian Dialect (TD) into MSA using a transformer model. Their model was trained on a corpus of 175K sentence pairs TD-MSA. Recently, [1] proceeded to translate the Omani dialect into English based on the transfer learning technique to adapt Marian NMT transformer [12]. Their model achieved a BLEU score of 9.88 on a test corpus of 87,200 posts.

Although this type of translation model which takes Arabic dialect as a source language gives acceptable results, it suffers from its inability to adapt to another different dialect. Therefore, to compensate for this insufficiency, we resort to multidialectal translation models. In this context, several researches have addressed this subject using transformers based on Large Language Models (LLMs). Certainly, LLMs have improved significantly the quality of machine translation, for this, researchers have resorted to evaluating their performance on various data, especially Arabic multidialectal parallel corpora. Reference [13] conducted a research to evaluate Bard (newly renamed Gemini) [9] and GPT-4 [20] on machine translation of 10 Arabic varieties (MSA and nine other Dialectal varieties) into English. The evaluation showed that GPT-4 outperformed Gemini on a manually parallel datasets with a gap of 2.35 points in BLEU scores. Reference [24] evaluated ChatGPT (ChatGPT-3.5 and ChatGPT-4) performance in translating 10 English health queries into Tunisian and Jordanian AD. These last two researches mentioned above highlighted a critical shortcoming of ChatGPT, GPT and Gemini in dialect translation, hence, the need to create robust linguistic model. Similarly, [27] concluded that the rapid evolve of the NLP fields requires creating more robust models and techniques in Arabic dialect identification and machine translation. They evaluated AraT5 (base), AraT5 (base-1024) [19], and AraBART [7] transformers to translate four AD (Egyptian, Emirati, Jordanian and Palestinian) into MSA based on the MADAR corpus [5]. Despite the merger of the three models, the overall BLEU score did not exceed 13.43 on the test set.

Reference [6] built a multidialectal translation model for four dialects into MSA based on the MADAR corpus [5]. This model, evaluated on 0-Shot learning (Emirati dialect) reached a significant BLEU score of 10.02. Reference [15] finetuned AraT5 transformer on four dialects (Egyptian, Emirati, Jordanian and Palestinian), and they achieved an overall BLEU on the Close dialect-MSA MT of 14.76. Reference [26] trained transformer-based models for MSA, Egyptian and Levantine dialects. They concluded that training the model on a corpus including different dialects greatly improves the translation of an unknown dialect. In addition, [2] proved the importance of using large-scale, dialect specific parallel corpora to reduce the effect of negative transfer from MSA caused by transformer-based NMT.

III. Datasets

Our approach aims to translate a message written in an Arabic dialect into MSA using a multidialectal translation system. For this, we treated two cases, the first focuses on the Levantine dialects, and the second concerns the Arab Maghreb dialects. For each case, we built parallel corpora for training and testing steps to fine-tune the transformers.

A. Levantine Datasets

We identified and made use of MADAR [5] and PADIC [18] parallel corpora from which we took a part of the Levantine-MSA sub-dataset. The latter was merged with our own data to build a Levantine parallel corpus containing 31114 entries of which 5000 are intended for testing and the rest for training the model. The data collected covered the Palestinian, Jordanian, and Syrian dialects as summarized in Table I.

TABLE I
LEVANTINE PARALLEL CORPUS

Dialect	Train entries	Test entries
Palestinian	3902	1000
Jordanian	6200	2000
Syrian	16012	2000
Total	26114	5000

B. Arab Maghreb Datasets

In addition to the new gathered data, we collected parallel data from the existing resources namely the augmented corpus "corpusNorthAfrica" of [6], the Multidialectal Parallel Corpus Arabic (MPCA) of [4] and PADIC [18]. Our final Maghreb parallel corpus covered the Tunisian, Moroccan, Algerian, and Libyan dialects. It was divided into 30000 entries for the training phase and 8000 entries to test the model as shown in Table II.

TABLE II
MAGHREB PARALLEL CORPUS

Dialect	Train entries	Test entries
Tunisian	8500	2000
Algerian	8500	2000
Moroccan	8500	2000
Libyan	4500	2000
Total	30000	8000

C. Arabic Dialect Corpus

As part of our contribution, we merged the two corpora mentioned above (Maghreb and Levantine corpora) in a single parallel corpus in order to use it in the evaluation phase.

TABLE III
ARABIC DIALECT CORPUS

Dialect	Train entries	Test entries
Tunisian	8500	2000
Algerian	8500	2000
Moroccan	8500	2000
Libyan	4500	2000
Palestinian	3902	1000
Jordanian	6200	2000
Syrian	16012	2000
Total	56114	13000

IV. FINE-TUNED TRANSFORMERS

Given the impressive results returned by transformers in the field of NLP and particularly in the neural machine translation task, our approach focused on Arabic multidialectal translation into MSA. To guarantee an attractive result, we fine-tuned various transformer models on the same datasets described above.

Our approach is based on two transfer transformers from the same family T5 which are T5X and the last updated version AraT5v2-base-1024. This choice is justified by the fact that T5X was considered an important achievement of the T5 family. T5X is very fast and it combines pre-training and fine-tuning to boost NLP performance but it was not pretrained on any Arabic dialect. AraT5v2-base-1024 transformer was trained on large and more diverse Arabic data including MSA and its dialects.

- *T5X*

T5X is a modular and high-performance open-source library for training, evaluating and inferring sequence models across many scales [23]. It is an improved version of the T5 text-to-text transfer transformer architectures provided by [22]. T5X can be used for pre-training from scratch or to fine-tune an existing language model. It was pre-trained on a large amount of data and can handle various tasks such as translation. The main optimization compared to T5 model is that it allows model and data parallelism. In fact, it supports GPU and TPU acceleration, and it was well-optimized for TPU.

- *AraT5*

AraT5 [19] is a pre-trained model from the T5 family. It is an optimization of the multilingual transfer transformer mT5 [28]. It focused on the Arabic varieties whether MSA or its dialects in various tasks including translation.

We chose the last update of the latest version AraT5v2-base-1024 which appeared just after the transformer AraT5-base. The feat of AraT5v2-base-1024 was that the sequence length was extended from 512 to 1024. Also, its last updated version was trained on large and more diverse data, and approximately 10 times faster than its predecessor.

The important thing in this study is that neither of the two models cited above was pre-trained on a parallel Arabic dialect-MSA corpus. Hence, the need to train from scratches each model on a large amount of data and adjusting the hyper-parameters to refine the quality of the returned results.

This research is designed to use PyTorch to fine-tune two transformer models namely T5X and AraT5v2-base-1024. To ensure a logical evaluation of our experiments, we used the same hyper-parameters for all fine-tuned models as detailed in Table IV. The parameter settings have been fixed with the aim of achieving the best performance. In fact, both for input (Arabic dialect) and target (MSA) data, the maximum length is set at 128 characters, and the batch size is fixed at 32. As long as the learning rate is crucial in the training phase and it ensures rapid and stable convergence of the model, we set its value at 2e-4 after several tests. In order to regularize the model and avoid overfitting, a value of 0.02 for the weight decay proved sufficient.

For each model, although the results increased significantly, the number of epochs was set to 10 due to execution resource constraints. Since the used Seq2SeqTrainer will save the model regularly and our dataset is quite large, we preferred to save three checkpoints maximum.

TABLE IV
PARAMETERS OF THE FINETUNED LLM MODELS

Parameter	Value
max __length	128
batch_size	32
learning_rate	2e-4
weight_decay	0.02
save_total_limit	3
num_train_epochs	10

V. EVALUATION AND RESULTS

All models were evaluated using the BiLingual Evaluation Understudy (BLEU) metric [21], which returns an approximate degree of similarity between a machine-translated text and a reference translation [15]. Based on the BLEU scores returned on test datasets, we noted that the last updated version of AraT5v2-base-1024 outperformed the T5X model by a significant margin. Table V summarizes the returned BLEU scores in the 10th epoch, for the three studied cases: Levantine-MSA, Maghrebin-MSA, and overall-MSA where overall signifies the fusion of the two first noted cases.

TABLE V
BLEU SCORES ON THE TEST DATASETS OF THE PROPOSED MODELS

Model	Training data		Test set BLEU
	Maghrebi-MSA	Levantine-MSA	
T5X	X	-	52.36
	-	X	29.15
	X	X	40.27
AraT5v2-base-1024	X	-	64.99
	-	X	48.38
	X	X	58.49

The evaluation was made only on the AraT5v2-base-1024 fine-tuned model for two reasons: the first is that it returned the best BLEU scores on the test dataset, and the second is that it belongs to the same family T5 as T5X and it is supposed to be the newest.

A. Zero-Shot Learning

Zero-Shot Learning (ZSL) is a technique where a model is tested on data on which it has never trained. In this context, and

in order to ensure the reliability of the best fine-tuned model (AraT5v2-base-1024), we tested it on a new dialect in each of the two studied groups of AD (Levantine and Maghrebi). For the Levantine dialect, we chose a test set of 2,000 sentences written in Lebanese dialect, while for the Maghrebin dialect, we tested the model on a Mauritanian dataset of 2,000 sentences. For the 0-shot, the model returned a low BLEU score of 21.93 while translating from Mauritanian into MSA, and a closely BLEU score of 42.51 when translating from Lebanese into MSA. As shown in Table VI, these scores influence on the global rates for the three studied cases.

TABLE VI
COMPARISON OF ARAT5V2-BASE-1024 AND THE ZSL

	Overall	Maghrebi-MSA	Levantine-MSA
AraT5v2-base-1024	58.49	64.99	48.38
ZSL	-	21.93	42.51

B. Interlingual Translation

We noticed that the transformers that we fine-tuned have not been pre-trained on parallel AD-MSA corpora, but they were based on X-MSA and AD-X parallel corpora where X represents a formal language. For this we opted for the back-translation technique where the English was used as a pivotal language to evaluate the model that returned the best results (AraT5v2-base-1024). In fact, the translation was done in two stages: first we translated from Arabic dialect into English, and then we translated from English into MSA.

As shown in Table VII, this interlingua approach failed to return significant result compared to the fine-tuned model AraT5v2-base-1024.

TABLE VII
COMPARISON OF ARAT5V2-BASE-1024 AND THE BACK-TRANSLATION

	Overall	Maghrebi-MSA	Levantine-MSA
AraT5v2-base-1024	58.49	64.99	48.38
Back-translation	52.46	61.22	43.15

C. Existing Tools

In our study, we leveraged the capabilities of Sider translator, a multilingual web-based tool designed recently for multiple languages including AD. This translation tool was built based on:

- ChatGPT-4 [20]: It is a large multimodal language model created by OpenAI. It was trained on a larger and more varied amount of data than its predecessor (ChatGPT-3.5) allowing it to provide more accurate and informative responses.
- Claude 2 [3]: It is an assistant created by Anthropic Company founded by former employees of OpenAI. It is an advanced language model trained using constitutional AI. It also uses the web to perform real-time searches, enabling it to provide up-to-date information.
- Gemini [9]: Formerly named Google Bard, it plays the role of an assistant during the process of translation based on a large language model developed by Google.

The choice of the Sider translator tool is justified by the fact that it encompasses the more recent and stronger LLMs and models (GPT-4, ChatGPT-4, Claude 2 and Gemini).

As shown in Table VIII, the obtained results reflect an enormous capability of fine-tuned AraT5v2-base-1024 model compared to Sider Translator in all cases despite a close BLEU score for the case of translation from Maghrebi dialects into MSA.

TABLE VIII
COMPARISON OF ARAT5V2-BASE-1024 AND SIDER TRANSLATOR

	Overall	Maghrebi-MSA	Levantine-MSA
AraT5v2-base-1024	58.49	64.99	48.38
Sider Translator	47.57	59.84	34.69

VI. DISCUSSION

Despite the emergence of several researches focusing on Arabic machine translation, there are few models focused on Arabic multidialectal translation into MSA. There is no doubt that LLM-based transformers have greatly and remarkably improved the quality of the retained results and especially for the monodialectal translation, but they require more optimization to adapt well to multilingual translation, and more precisely having perfect multidialectal translation model into MSA.

The translation results of any model depend on the quality and size of the data, hence the obligation to carefully select the parallel training corpora and test datasets. In response, we propose a framework which showed its effectiveness in addressing the intricate challenges posed by Arabic multidialectal translation into MSA. Two large regional dialectal groups were considered in this work, namely Levantine dialects (Palestinian, Jordanian, and Syrian) and Maghrebi dialects (Tunisian, Algerian, Moroccan, and Libyan). We justify our choice by the fact that all researches conducted on monodialectal translation into MSA failed to retain good results when applied on other dialects. Although there are major differences between AD, we note a certain resemblance at the level of each region (Maghrebi, Levantine, Gulf, and Egyptian) [11].

The parallel corpora to be prepared must respect the nature and settings of the model to be finetuned. In this context, we chose two models from the T5 family: T5X and the last updated version of AraT5v2-base-1024. Unlike the second model, the first was not pertained on AD. Therefore, AraT5v2-base-1024 achieved acceptable results, while T5X failed given the lack of sufficient data. For this reason, we tried to diversify the nature and sources of parallel data. We gathered data from various resources including those not used during the pretraining phase of the models.

Another contribution of this study is the corpora size which directly influences the achieved scores. We remarked that the T5 family models (even the last version of AraT5v2-base-1024) were not pretrained on AD-MSA parallel corpora. Therefore, enlarging the datasets clearly improved the BLEU scores. For the two multidialectal translation tasks, Levantine-MSA and Maghrebi-MSA, AraT5v2-base-1024 outperforms T5X transfer transformer and returned BLEU scores of 48.38 and 64.99, respectively.

We evaluated the effectiveness of AraT5v2-base-1024 in

three cases. First, we fine-tuned the model on a single dataset by merging the Levantine-MSA and the Maghrebi-MSA parallel corpora. The model retained its capability and achieved an overall BLEU score of 58.49. In addition, we conducted a ZSL on each of the two region dialects using in each case a test set of 2,000 sentences. On Mauritanian dialect, it returned a low BLEU score of 21.93 due to the mixture of dialects (French, Hassaniya, etc.). This low score reflects its degree of similarity with Maghrebi dialects. While on Lebanese dialect, which is very close to the Levantine dialects, the achieved BLEU score was 42.51.

Second, we focused on the fact that the fine-tuned model was pretrained on parallel AD-English and English-MSA corpora. For this, we carried out a multidialectal translation into MSA using English as a pivotal language. The translation process was done in two stages: Arabic multidialectal translation into English, then translation from English into MSA. Unfortunately, the model failed to exceed the BLEU scores achieved in the experiment phase, and it returned BLEU scores of 43.15, 61.22, and 52.46 against 48.38, 64.99, and 58.49, for Levantine-MSA, Maghrebi-MSA, and Overall-MSA, respectively.

The third evaluation was an attempt to compare our model with open-source tools. We chose Sider Translator tool, which is based on three among the recent and strong LLMs and models (GPT-4, ChatGPT-4, Claude 2 and Gemini). Successively, Sider returned BLEU scores of 34.69, 59.84, and 47.57 against 48.38, 64.99, and 58.49 for AraT5v2-base-1024, for Levantine-MSA, Maghrebi-MSA, and Overall-MSA, respectively.

This outperformance is due to the fact that the transformers are based on the multihead self-attention mechanism which allows the models to learn better, and therefore return better results. Moreover, we believe that our choice of varied and consistent training data strongly influences the quality of the retained results.

VII. Conclusion and Future Work

Besides the lack of parallel resources which hinders the ability to effectively process Arabic multidialectal machine translation into MSA based on recent transfer transformers, particularly AraT5v2-base-1024, our study showed the sufficiency of these models against the back-translation and the open-source tools. Our efforts yielded notable strengths. Firstly, the grouping of AD by region has remarkably improved the quality of the obtained results. Additionally, a good choice of data, whether in terms of dialects or in quality and quantity of data, clearly influences on the retained results. Moreover, the parameter settings of the model are so important to increase the evaluation metric value.

For future directions, an extension of additional AD will be considered to expand the ability of the model to translate any Arabic dialect into MSA. Additionally, these attractive results encourage us to move on to the understanding stage.

References

[1] Khoula Al-Kharusi and Abdurahman Abdulsalam. 2023. Machine Translation of Omani Arabic Dialect from Social Media. In *Proceedings of ArabicNLP 2023*, 2023. Association for Computational Linguistics, Singapore (Hybrid), 302-309. https://doi.org/10.18653/v1/2023.arabicnlp-1.24

[2] Rania Al-Sabbagh. 2023. The Negative Transfer Effect on the Neural Machine Translation of Egyptian Arabic Adjuncts into English: The Case of Google Translate. *IJAES* (October 2023). https://doi.org/10.33806/ijaes.v24i1.560

[3] Anthropic. 2023. Claude 2. Retrieved from https://www.anthropic.com/news/claude-2

[4] Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014* (2014), 1240-1245.

[5] 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. (2018).

[6] 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*, 2023. Association for Computational Linguistics, Singapore (Hybrid), 683-689. https://doi.org/10.18653/v1/2023.arabicnlp-1.75

[7] Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization. Retrieved April 16, 2024 from http://arxiv.org/abs/2203.10945

[8] Alexander Erdmann, Nizar Habash, Dima Taji, and Houda Bouamor. 2017. Low Resourced Machine Translation via Morpho-syntactic Modeling: The Case of Dialectal Arabic. Retrieved September 21, 2023 from http://arxiv.org/abs/1712.06273

[9] Google. 2023. Gemini (BARD). Retrieved from https://gemini.google.com/app

[10] Ebtesam H Almansor and Ahmed Al-Ani. 2017. Translating Dialectal Arabic as Low Resource Language using Word Embedding. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, November 10, 2017. Incoma Ltd. Shoumen, Bulgaria, 52-57. https://doi.org/10.26615/978-954-452-049-6_008

[11] Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02139-8

[12] 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*, 2018. Association for Computational Linguistics, Melbourne, Australia, 116-121. https://doi.org/10.18653/v1/P18-4020

[13] Karima Kadaoui, Samar M Magdy, Abdul Waheed, Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023: Evaluation of Bard and ChatGPT on Machine Translation of Ten Arabic Varieties. Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023), (2023), pages 52-75.

[14] 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 (March 2023), 1-21. https://doi.org/10.1145/3568674

[15] Abdullah Khered, Ingy Yasser 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 Dialectal Text to Modern Standard Arabic. Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), (December 2023), pages 658-664.

[16] Philipp Koehn, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, and Christine Moran. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07*, 2007. Association for Computational Linguistics, Prague, Czech Republic, 177. https://doi.org/10.3115/1557769.1557821

[17] Mohamed Lichouri and Mourad Abbas. 2021. Machine Translation for Zero and Low-resourced Dialects using a New Extended Version of the

Dialectal Parallel Corpus (Padic v2.0). *Proceedings of the 4th International Conference on Natural Language and Speech Processing* 4th, (2021).

[18] K Meftouh, S Harrat, and Kamel Smaïli. 2018. PADIC: extension and new experiments. *International Conference on Advanced Technologies ICAT* 7th, (April 2018).

[19] 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)*, 2022. Association for Computational Linguistics, Dublin, Ireland, 628-647. https://doi.org/10.18653/v1/2022.acl-long.47

[20] OpenAI. 2023. GPT-4. Retrieved from https://openai.com/research/gpt-4

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001. Association for Computational Linguistics, Philadelphia, Pennsylvania, 311. https://doi.org/10.3115/1073083.1073135

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Retrieved April 9, 2024 from http://arxiv.org/abs/1910.10683

[23] Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling Up Models and Data with t5x and seqio. Retrieved April 9, 2024 from http://arxiv.org/abs/2203.17189

[24] Malik Sallam and Dhia Mousa. 2024. Evaluating ChatGPT performance in Arabic dialects: A comparative study showing defects in responding to Jordanian and Tunisian general health prompts. *MJAIH* 2024, (January 2024), 1-7. https://doi.org/10.58496/MJAIH/2024/001

[25] Wael Salloum and Nizar Habash. 2012. Elissa: A Dialectal to Standard Arabic Machine Translation System. Proceedings of COLING 2012: Demonstration Papers, (December 2012), pages 385-392.

[26] Pamela Shapiro and Kevin Duh. 2019. Comparing Pipelined and Integrated Approaches to Dialectal Arabic Neural Machine Translation. *Association for Computational Linguistics* Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, (June 2019), 214-222. https://doi.org/10.18653/v1/W19-1424

[27] Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. DialectNLU at NADI 2023 Shared Task: Transformer Based Multitask Approach Jointly Integrating Dialect and Machine Translation Tasks in Arabic. In *Proceedings of ArabicNLP 2023*, 2023. Association for Computational Linguistics, Singapore (Hybrid), 614-619. https://doi.org/10.18653/v1/2023.arabicnlp-1.63

[28] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. Retrieved April 9, 2024 from http://arxiv.org/abs/2010.11934