UniManc at NADI 2023 Shared Task: A Comparison of Various T5-based Models for Translating Arabic Dialectal Text to Modern Standard Arabic

Abdullah Khered[1,2] , **Ingy Yasser Abdelhalim** [1], **Nadine Abdelhalim**[1],
Ahmed Soliman [1] and **Riza Batista-Navarro**[1]
[1]The University of Manchester, UK [2]King Abdulaziz University, Saudi Arabia
abdullah.khered@manchester.ac.uk, ingyyh@live.com, nadineyh@live.com,
ahmedsoliman360@gmail.com, riza.batista@manchester.ac.uk

Abstract

This paper presents the methods we developed for the Nuanced Arabic Dialect Identification (NADI) 2023 shared task, specifically targeting the two subtasks focussed on sentence-level machine translation (MT) of text written in any of four Arabic dialects (Egyptian, Emirati, Jordanian and Palestinian) to Modern Standard Arabic (MSA). Our team, UniManc, employed models based on T5: multilingual T5 (mT5), multi-task fine-tuned mT5 (mT0) and AraT5. These models were trained based on two configurations: joint model training for all regional dialects (J-R) and independent model training for every regional dialect (I-R). Based on the results of the official NADI 2023 evaluation, our I-R AraT5 model obtained an overall BLEU score of 14.76, ranking first in the Closed Dialect-to-MSA MT subtask. Moreover, in the Open Dialect-to-MSA MT subtask, our J-R AraT5 model also ranked first, obtaining an overall BLEU score of 21.10.

1 Introduction

The Arabic language serves as a linguistic umbrella for approximately 420 million speakers, predominantly dispersed across 22 countries in the Middle East and North Africa (MENA) region. A defining characteristic of the language is its diglossic nature, where Modern Standard Arabic (MSA) coexists with a myriad of dialects, commonly referred to as Dialectal Arabic (DA) (Al-Sobh et al., 2015; Abdul-Mageed et al., 2022).

MSA is the formal version of the Arabic language, employed in educational settings, official documents and written literature. It serves as a standardised communication medium across Arabic-speaking countries. In contrast, DA represents the colloquial forms of Arabic, which are more region-specific and employed in day-to-day verbal interactions (Shoufan and Alameri, 2015). Notably, dialects can vary significantly based on geographic location and socio-economic factors, ranging from subtle differences to being nearly mutually unintelligible. This linguistic variation presents considerable challenges for machine translation (MT) models trained on MSA. These models often fail to capture the nuanced differences in dialects, resulting in poor translation performance when applied to DA. Compounding this issue is the scarcity of parallel corpora containing MSA translations of text written in DA, limiting resources for model training and evaluation (Harrat et al., 2019).

In the context of these challenges, this paper aims to explore the extent to which various sequence-to-sequence models based on the Text-to-Text Transfer Transformer, popularly known as T5 (Raffel et al., 2020), can translate a source text written in an Arabic dialect to a target text that is written in MSA. We participated in the Nuanced Arabic Dialect Identification (NADI) 2023 Shared Task (Abdul-Mageed et al., 2023), specifically in Subtasks 2 and 3, described below.

Subtask 2: Dialect-to-MSA MT - Closed Task. The objective of this subtask is sentence-level machine translation from four dialects (Egyptian, Emirati, Jordanian and Palestinian) to MSA. Participants were restricted to using the MADAR parallel corpus (Bouamor et al., 2019) for training and were asked to evaluate their models on newly released development and test sets.

Subtask 3: Dialect-to-MSA MT - Open Task. This subtask is similar to Subtask 2, except for the fact that participants were allowed to utilise additional datasets for model training. One of the goals of this subtask is to encourage the creation of new parallel corpora to facilitate future research.

Apart from investigating the performance of various T5-based models on the above-mentioned tasks, our work makes an additional contribution by developing a new dataset, Emi-NADI, which contains MSA translations of sentences written in Emirati, one of the most under-resourced dialects.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 658–664
December 7, 2023 ©2023 Association for Computational Linguistics

The Emi-NADI dataset and the code for developing and evaluating our models for Subtasks 2 and 3 have been made publicly available.[1]

2 Datasets

This section describes the datasets that were utilised in training our models.

2.1 The MADAR Corpus

As mentioned in the previous section, participants in the closed version of the dialect-to-MSA translation task, Subtask 2, were allowed to use only the MADAR parallel corpus (Bouamor et al., 2019), which covers the dialects used in 25 Arabic-speaking cities, as well as English and MSA.

2.2 Additional Corpora

In the open version of the dialect-to-MSA machine translation task, Subtask 3, participants were allowed to leverage any dataset. As we searched for potentially useful publicly available datasets, we considered those that cover various Arabic dialects, including regional ones that are relevant to the four countries of interest in NADI. For example, the Gulf dialect is relevant to Emirati (since the United Arab Emirates is one of the Gulf countries), and the Levantine dialect is relevant to Jordanian and Palestinian (since Jordan and Palestine belong to the Levant). Apart from the MADAR corpus, we identified and made use of four datasets: (1) PADIC, (2) Dial2MSA, (3) a semantic textual similarity (STS) dataset for Arabic dialects, and (4) our own Emi-NADI dataset containing Emirati-to-MSA translations. Table 1 provides information on the size of each dataset in terms of number of dialectal sentences with translations to MSA.

Dataset	Egy.	Gulf	Lev.
MADAR	13,800	15,400	18,600
PADIC	0	0	12,824
Dial2MSA	16,355	0	0
Arabic STS	2,758	2,758	0
Emi-NADI	0	2,712	0
Total	32,913	20,870	31,424

Table 1: The number of dialect-to-MSA translation pairs in each of the datasets used in Subtask 3.

PADIC (Meftouh et al., 2018) is a parallel corpus containing dialectal Arabic texts covering six Arab cities including Gaza and Damascus, which are both in the Levant region. Meanwhile, the Dial2MSA dataset (Mubarak, 2018) consists of tweets written in four Arabic dialects (Egyptian, Gulf, Levantine, Maghrebi) and their corresponding MSA translations. As only the translations for Egyptian and Maghrebi were manually validated, we made use of the Egyptian-to-MSA translations only. In the work of Al Sulaiman et al. (2022) that focussed on Arabic STS (i.e., determining the semantic similarity between two given sentences), they manually produced MSA, Egyptian and Saudi dialect translations for 2758 English sentences, which we also utilised in our work.

Our own dataset, Emi-NADI, was created to address the scarcity of parallel corpora covering the Emirati dialect, and contains MSA translations of the Emirati tweets in the training datasets provided as part of NADI Subtask 1 (country-level dialect identification) (Abdul-Mageed et al., 2020, 2021, 2023). The translations were generated by a large language model (LLM), specifically the GPT 3.5 Turbo model,[2] resulting in a total of 2712 translations. A subset of 1000 automatically generated translations were manually validated (by native Arabic speakers who understand Emirati) to ensure quality. Both the validated and the non-validated samples were used in model training.

3 Methodology

In this section, we introduce the T5-based models that we built upon, explain how they were fine-tuned and discuss hyperparameter optimisation.

3.1 Models

T5 casts different natural language processing (NLP) tasks into a standard text-to-text format. One of the NLP tasks that T5 was already trained on is machine translation (Raffel et al., 2020). In this work, we fine-tuned three types of T5 models, namely, AraT5, mT5 and mT0.

AraT5. AraT5 (Nagoudi et al., 2022) is based on the same architectural foundation as the original T5 models, but trained specifically on Arabic data encompassing both MSA and dialectal Arabic (tweets). The most recent version of AraT5, AraT5$_{\text{v2}}$,[3] was used in all our experiments.

Multilingual T5 (mT5). mT5 (Xue et al., 2021) is a multilingual variant of T5 that underwent pre-

[1] https://github.com/khered20/UniManc_NADI2023_ArabicDialectToMSA_MT

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://huggingface.co/UBC-NLP/AraT5v2-base-1024

Source	Target
Original Pair	
الجرح بتاعي بيألم	جرحي يؤلمني
Additional Pair	
جرحي يؤلمني	جرحي يؤلمني

Table 2: An example of the additional training pair where each of the source and target is the text written in MSA (English translation: *"My wound hurts"*). The tokens shown in grey in the Egyptian source text of the original pair share the same root as the tokens in grey in the target MSA text.

training using a novel dataset sourced from Common Crawl, encompassing 101 languages. Although its pre-training process is underpinned by the original T5 architecture, it incorporated some improvements, such as the adoption of a different activation function in the feed-forward layer (i.e., GeGLU instead of the conventional RELU).

Multi-task fine-tuned mT5 (mT0). Multitask-prompted fine-tuning (MTF) has demonstrated its efficacy in assisting LLMs in adapting to novel tasks within a zero-shot setting. In this vein, mT0 is a multitask-prompted fine-tuned version of mT5. mT0 has showcased remarkable zero-shot generalisation capabilities, even when presented with languages it has never encountered before (Muennighoff et al., 2023).

3.2 Training Configurations

In the early stages of our experimentation, we noticed that many dialectal texts contain words that are shared between a dialect and MSA. Thus, for every translation pair in our training data, we generated an additional pair where each of the source and target is the text written in MSA. An example is provided in Table 2. Our models were then trained — based on the two different configurations outlined below — using these additional pairs, enabling them to learn how to handle sentences that include words that are also used in MSA.

Training a joint model for all regional dialects (J-R). In this configuration, all dialect-to-MSA translation pairs (in the training sets for Subtasks 2 and 3) that correspond to the regions relevant to the four dialects of interest were utilised in training one model. Therefore, translation pairs from datasets that cover the Egyptian, Gulf and Levantine dialects were utilised in model training. The result is one joint model trained to translate dialectal text to MSA, regardless of which dialect it was written in.

Training an independent model for each regional dialect (I-R). In this configuration, one model was trained for every relevant regional dialect. This resulted in four separate models, where each model was independently trained to translate texts written in one specific dialect only, to MSA.

3.3 Hyperparameter Optimisation

For each of the two subtasks, we trained our models using two Nvidia A100 GPUs based on the configurations described above. All models accept input sequences with a maximum length of 128 tokens and generate output text also with a maximum length of 128 tokens. Learning rate and batch size were fixed at 5e-5 and 16, respectively. The maximum number of epochs was set to 40, although we always selected the model produced in the epoch that yielded the best performance on the development (dev) set provided by the NADI organisers. Importantly, we investigated whether incorporating beam search (Freitag and Al-Onaizan, 2017) during translation leads to improved performance, experimenting with different beam sizes ranging from 1 to 5.

4 Evaluation and Results

All models for Subtasks 2 and 3 were evaluated using the BiLingual Evaluation Understudy (BLEU) metric (Papineni et al., 2002), which estimates the similarity between a machine-translated text and a reference translation based on overlapping tokens.

The results of our joint regional (J-R) and independent regional (I-R) models for Subtasks 2 and 3, without using beam search (i.e., beam size = 1), are shown in Tables 3 and 4, respectively. One can observe in Table 3 that for Subtask 2, in all cases (except for Jordanian), the I-R version of a model consistently outperforms its J-R counterpart. This finding led us to further experiment with the I-R models by investigating different values for beam size. The results, shown in Table 7 in the Appendix, helped us in identifying the best-performing I-R models. Based on this, we selected two I-R AraT5 models, one I-R mT5 model and one I-R mT0 model to comprise our set of models for the official evaluation (on the NADI test set), together with the best J-R model.

Model	Egy.	Emi.	Jor.	Pal.	Overall
Joint Regional Models (J-R)					
mT0	12.28	**10.98**	10.06	9.65	11.12
mT5	12.16	10.93	9.14	9.49	11.13
AraT5$_{v2}$	**14.65**	10.65	**11.20**	**10.53**	**13.30**
Independent Regional Models (I-R)					
mT0	13.88	12.91	9.55	10.91	12.53
mT5	15.02	**15.25**	10.32	10.69	13.57
AraT5$_{v2}$	**17.21**	14.13	**12.14**	**13.33**	**15.14**

Table 3: Comparison of joint regional (J-R) and independent regional (I-R) models for Subtask 2, based on the development set. Beam size = 1.

Model	Egy.	Emi.	Jor.	Pal.	Overall
Joint Regional Models (J-R)					
mT0	15.11	26.85	18.44	15.16	18.25
mT5	18.80	29.04	18.63	15.50	19.81
AraT5$_{v2}$	**20.23**	**32.84**	**24.85**	**18.27**	**23.37**
Independent Regional Models (I-R)					
mT0	18.28	27.35	19.82	16.46	19.96
mT5	18.26	26.83	21.45	16.48	20.25
AraT5$_{v2}$	**21.90**	**31.28**	**24.45**	**18.08**	**23.45**

Table 4: Comparison of joint regional (J-R) and independent regional (I-R) models for Subtask 3, based on the development set. Beam size = 1.

In the comparison of the J-R and I-R models (without beam search) for Subtask 3 shown in Table 4, it is evident that the AraT5 models outperform both mT0 and mT5 by a noticeable margin, and that the I-R models outperform their J-R counterparts overall. We thus further experimented with the I-R versions of the AraT5 model by investigating different beam sizes. The results, shown in Table 8 in the Appendix, informed our selection of models for the official evaluation (on the NADI test set), which consists of the three best I-R AraT5 models, one I-R mT5 model and the best J-R model.

Tables 5 and 6 present the results of our chosen models on the NADI test sets for Subtasks 2 and 3, respectively. As shown in Table 5, the I-R AraT5 model with beam size = 3 outperformed our other models (obtaining a score of 14.76). Meanwhile, our Subtask 3 results, shown in Table 6, demonstrate that the J-R AraT5 model (with beam size = 1) performs best overall (21.10). To investigate whether adjusting the beam size of the J-R AraT5 model will lead to even better performance, we submitted the same model to the post-evaluation phase of Subtask 3, but this time with beam size = 5. The overall score did increase to 21.87, implying once again that incorporating beam search leads to better performance.

5 Discussion

In Tables 3 and 4, it can be observed that for both subtasks the independent regional (I-R) models performed better compared to the joint regional (J-R) models, with AraT5 performing the best overall. This can be explained by the fact that AraT5 was trained with a specific focus on Arabic whereas the others (mT0 and mT5) were trained on many other languages apart from Arabic. This implies that for the dialect-to-MSA translation task, a model that was trained solely on the Arabic language is superior over multilingual models.

Given that the I-R models performed better, multiple beam sizes were explored. Our results show that increasing the beam size leads to an improvement in overall performance. However, it is worth noting that the optimal beam size could vary between the development and test sets (e.g., beam size = 4 on the development set and beam size = 3 on the test set for Subtask 2), although the difference in performance is very marginal.

Error analysis was conducted to qualitatively evaluate our best-performing model for Subtask 3. Specifically, we analysed cases where the model obtained low BLEU scores and manually assessed the quality of the translations produced by the model. An example for each dialect is shown in Table 10 in the Appendix. Interestingly, the model's translations of the Egyptian, Emirati and Jordanian source texts are arguably correct, as they convey the same meaning as the reference translations. They, however, obtained low BLEU scores due to the fact that the BLEU metric takes into account lexical but not semantic similarity, in comparing a generated translation with a reference one. As for the Palestinian example, the model's failed translation can be attributed to code-mixing, i.e., the presence of the non-Arabic word "bravo" (written in Arabic script) in the source text.

6 Conclusion and Future Work

In this paper, we describe the approaches we developed for NADI 2023 Subtask 2 (Closed Dialect-to-MSA MT) and Subtask 3 (Open Dialect-to-MSA MT). Our results reveal that fine-tuning AraT5 and incorporating beam search during translation lead to top-ranking performance. Possible future directions include the development of a multilingual model focussed on Arabic dialects and MSA, and the creation of further parallel corpora covering low-resourced Arabic dialects.

Model	Configuration	Beam	Egy.	Emi.	Jor.	Pal.	Overall
AraT5$_{v2}$	J-R	1	12.50	10.15	11.39	10.28	12.12
mT0	I-R	3	13.64	12.43	7.67	9.32	11.37
mT5	I-R	2	14.04	10.42	10.65	11.66	12.38
AraT5$_{v2}$	I-R	3	16.04	**14.30**	12.55	**13.55**	**14.76**
AraT5$_{v2}$	I-R	4	**16.54**	14.20	**12.73**	13.04	14.73

Table 5: Results of evaluating our submitted models on the NADI Subtask 2 test set.

Model	Configuration	Beam	Egy.	Emi.	Jor.	Pal.	Overall
AraT5$_{v2}$	J-R	1	17.65	**28.46**	**22.03**	17.29	**21.10**
mT5	I-R	1	15.75	25.15	16.44	16.15	17.95
AraT5$_{v2}$	I-R	1	17.95	24.94	20.84	17.67	20.22
AraT5$_{v2}$	I-R	3	19.61	25.79	20.95	**18.31**	21.02
AraT5$_{v2}$	I-R	4	**19.70**	26.02	21.00	18.27	21.08

Table 6: Results of evaluating our submitted models on the NADI Subtask 3 test set.

Limitations

Due to time and computational resource constraints, we were unable to conduct a more systematic investigation of the effect of different beam size values for the joint regional AraT5, mT5 and mT0 models that we employed.

Furthermore, most of the models that we submitted to the official NADI 2023 Subtasks 2 and 3 evaluation were trained following a configuration whereby a separate model was independently trained on every dialect. This means that prior to translation, the dialect in which an input text was written in needs to be predetermined, so that the relevant model can be applied.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, ElMoatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

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. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. NADI 2022: The third nuanced Arabic dialect identification shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mahmoud Al-Sobh, Abdel-Rahman Abu-Melhim, and Nedal Bani Hani. 2015. Diglossia as a result of language variation in arabic: Possible solutions in light of language planning. *Journal of Language Teaching and Research*, 6:274.

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):1–14.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. 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.

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

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing Management*, 56(2):262–273. Advance Arabic Natural Language Processing (ANLP) and its Applications.

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. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), OSACT2018 Workshop*, pages 49–53.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, and Teven et al. Le Scao. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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 ACL*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Appendix

Model	Beam size	Egy.	Emi.	Jor.	Pal.	Overall
AraT5$_{v2}$	1	17.209	14.127	12.143	13.329	15.139
	2	17.152	15.197	12.906	14.458	15.828
	3	18.702	14.798	12.892	14.507	**16.166**
	4	19.092	15.281	12.478	14.552	**16.173**
	5	19.274	15.052	12.213	14.267	16.037
mT5	1	15.023	15.253	10.324	10.689	13.57
	2	15.755	14.888	11.924	10.345	**13.919**
	3	16.121	14.903	11.395	9.962	13.757
	4	16.076	14.857	11.754	10.015	13.873
	5	16.071	14.744	11.519	10.205	13.909
mT0	1	13.882	12.914	9.551	10.907	12.525
	2	13.199	12.371	10.398	10.884	12.389
	3	14.498	12.336	11.198	11.692	**13.222**
	4	14.432	12.69	10.643	11.633	13.085
	5	14.439	12.554	10.283	11.429	12.989

Table 7: Results of using different values for beam size on the Subtask 2 development set, obtained by independent regional (I-R) models. The best performing models (whose overall scores are shown in bold) formed the basis of our submission to the official NADI 2023 evaluation.

Model	Beam size	Egy.	Emi.	Jor.	Pal.	Overall
AraT5$_{v2}$	1	21.897	31.28	24.451	18.077	**23.452**
	2	22.153	30.65	23.922	17.628	23.151
	3	22.788	31.123	24.035	17.525	**23.447**
	4	22.59	31.174	23.763	17.559	**23.334**
	5	22.401	31.291	23.915	17.56	23.319

Table 8: Results of using different values for beam size on the Subtask 3 development set, obtained by independent regional (I-R) versions of the AraT5 model. The best performing models (whose overall scores are shown in bold) formed the basis of our submission to the official NADI 2023 evaluation.

Dialect	Dialectal Text	English Translation
Egy	لكن التأمين أنا مبفهمش فيها خالص.	But the insurance I don't understand it at all
Emi	إنزين و اللولو جم يسوى؟	Then how much is this pearl worth?
Jor	يسلموا ايديكي عالفطور الطيب	Thank you for the delicious breakfast
Pal	برافو يا شفيق ترابايتي	Bravo Shafiq, my upbringing

Table 9: English translations of the dialectal Arabic examples shown in Table 10, provided for the reader's convenience.

Dialect	Dialectal Text	Reference MSA Translation	Model Translation
Egy	لكن التأمين أنا مبفهمش فيها خالص.	لكن التأمين ، أنا لا أفهم فيه أبداً.	لكن التأمين لا أفهم فيه على الإطلاق.
Emi	إنزين و اللولو جم يسوى؟	حسنًا و اللؤلؤ كم يساوى؟	ما قيمة هذا اللولو ؟
Jor	يسلموا ايديكي عالفطور الطيب	سلمت يداكِ على هذا الفطور اللذيذ	شكراً جزيلاً على الإفطار الطيب.
Pal	برافو يا شفيق ترابايتي	أحسنت يا شفيق، تعليمي	برافو يا شفيق تربيتي

Table 10: Examples showing cases where the translation generated by our best-performing Subtask 3 model was given a low BLEU score despite being semantically correct. For English translations of the dialectal examples, we refer the reader to Table 9.