AraT5: Text-to-Text Transformers for Arabic Language Generation

El Moatez Billah Nagoudi[⋆] **AbdelRahim Elmadany**[⋆] **Muhammad Abdul-Mageed**[⋆]
Deep Learning and Natural Language Processing Group
The University of British Columbia
{moatez.nagoudi,a.elmadany,muhammad.mageed}@ubc.ca

Abstract

Transfer learning with a unified Transformer framework (T5) that converts all language problems into a text-to-text format was recently proposed as a simple and effective transfer learning approach. Although a multilingual version of the T5 model (mT5) was also introduced, it is not clear how well it can fare on non-English tasks involving *diverse* data. To investigate this question, we apply mT5 on a language with a wide variety of dialects–Arabic. For evaluation, we introduce a novel benchmark for **AR**abic language **GEN**eration (ARGEN), covering *seven* important tasks. For model comparison, we pre-train three powerful Arabic T5-style models and evaluate them on ARGEN. Although pre-trained with $\sim$ 49% less data, our new models perform significantly better than mT5 on *all* ARGEN tasks (in 52 out of 59 test sets) and set several new SOTAs. Our models also establish new SOTA on the recently-proposed, large Arabic language understanding evaluation benchmark ARLUE (Abdul-Mageed et al., 2021). Our models are publicly available. We also link to individual ARGEN datasets through our public repository.[1]

1 Introduction

Due to their remarkable ability to transfer knowledge from unlabeled data to downstream tasks, pre-trained Transformer-based language models have emerged as important components of modern natural language processing (NLP) systems. In particular, the unified framework that converts all text-based language problems into a text-to-text format presented through the T5 model (Raffel et al., 2019) is attractive. In addition to its simplicity, this approach is effective since it allows knowledge transfer from high-resource to low-resource tasks without the need for changing model architecture. Unlike models such as BERT (Devlin et al., 2019), which are based on encoders only, the T5 model is an encoder-decoder that can naturally be employed for natural language generation. Although the T5 model, originally pre-trained for English, was recently extended to the multilingual setting as mT5 (Xue et al., 2020), it is not clear how suited it is to individual languages (and varieties of these languages). In addition, systematic issues have been discovered in multilingual corpora on which language models have been trained (Kreutzer et al., 2021). In absence of comparisons with monolingual pre-trained language models that serve different non-English contexts, it remains unknown how multilingual models really fare against language-specific models.

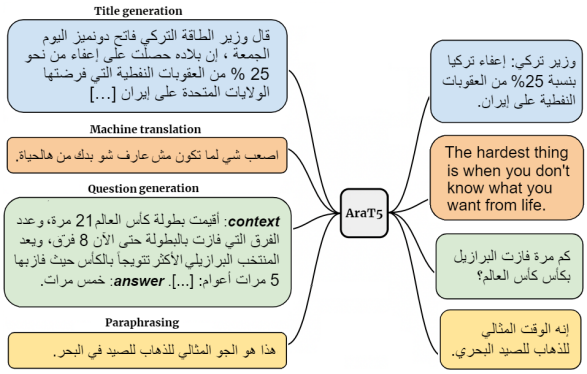

Figure 1: Our AraT5 encoder-decoder model and prompt samples from four investigated tasks, namely: title generation, machine translation, question generation, and paraphrasing.

In this work, we offer the first comparison of the mT5 model to similar encoder-decoder models dedicated to Arabic. We choose Arabic as our context due to its large set of diverse varieties as well as its wide use on social media. Our work aims at uncovering the extent to which mT5 can serve Arabic's different varieties. Our work also meets an existing need for pre-trained Transformer-based sequence-to-sequence models. In other words, while several BERT-based models have been pre-trained for Arabic (Antoun et al., 2020; Abdul-Mageed et al.,

[1] https://github.com/UBC-NLP/araT5
[⋆] All authors contributed equally.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 628 - 647

2021; Inoue et al., 2021), no such attempts have been made to create sequence-to-sequence models that we know of. Another motivation for our work is absence of an evaluation benchmark for Arabic language generation tasks. Apart from machine translation where researchers are starting to propose benchmarks such as AraBench (Sajjad et al., 2020), there are no benchmarks that can be used to methodically measure Arabic natural language generation performance.

Our main contributions are as follows: **(1)** We introduce three powerful variants of the text-to-text transformer (T5) model dedicated to Modern Standard Arabic (MSA) and a diverse set of Arabic dialects. We include in our vocabulary 11 languages other than Arabic (e.g., English, French, German, Russian), which also allows us to evaluate our models under zero-shot pre-training conditions involving these languages. **(2)** We propose a novel unified benchmark for **AR**abic natural language **GE**neration (**ARGEN**) composed of *seven* tasks: machine translation, code-switched text translation, summarization, news title generation, question generation, paraphrasing, and transliteration. ARGEN is collected from a total of 19 datasets, including 9 *new datasets* proposed in this work. **(3)** To show the utility of our new models, we evaluate them on ARGEN under both *full* and *zero-shot* pre-training conditions. Our models set new SOTA on the majority of datasets in *all* seven tasks. **(4)** Although the main focus of our work is language *generation*, we also show the effectiveness of our models on Arabic language *understanding* by fine-tuning our new models on a large, recently proposed Arabic language understanding benchmark. Again, our models establish new SOTA on the majority of language understanding tasks.

The rest of the paper is organized as follows: Section 2 describes our Arabic pre-tained models. In Section 3, we introduce ARGEN, our new natural language generation benchmark. We evaluate our models on ARGEN in Section 4. Section 5 is an analysis and discussion of our results. In Section 6, we provide an overview of related work. We conclude in Section 7. We now introduce our new pre-trained models.

2 Our Models

2.1 Pre-Training Data

MSA Data. We use 70GB of MSA text (7.1B tokens) from the following sources: AraNews (Nagoudi et al., 2020), El-Khair El-Khair (2016), Gigaword,[2], OSCAR (Suárez et al., 2019), OSIAN (Zeroual et al., 2019), Wikipedia Arabic, and Hindawi Books.[3]

Twitter Data. We randomly sample 1.5B Arabic tweets (178GB) from a large in-house dataset of $\sim$ 10B tweets. We use string matching to only include tweets with at least 3 Arabic words, regardless whether the tweet has non-Arabic string or not.

Our combined MSA and Twitter data make up 29B tokens, and hence is $\sim$ 49% less than Arabic tokens on which mT5 is pre-trained (57B Arabic tokens). More information about our pre-training data is in Table 1.

MSA Vs. Dialect Distribution. In order to analyze MSA-dialect distribution in our Twitter data, we run the binary (MSA-dialect) classifier introduced in Abdul-Mageed et al. (2020b) on a random sample of 100M tweets. We find the data to involve 28.39% predicted dialect tweets and 71.61% predicted MSA. We also acquire country-level dialect labels using an in-house strong classifier on the dialectal portion of the data (i.e., $\sim$ 28.39 millions tweets), finding dialectal tweets to be truly *geographically diverse* as shown in Figure 2.

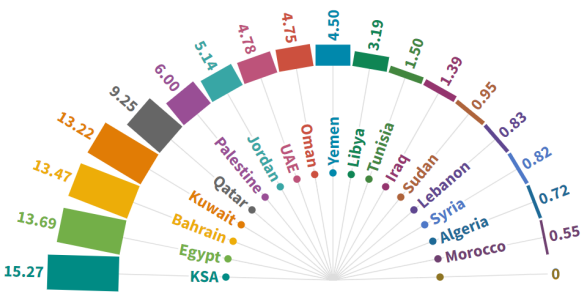

Figure 2: Country-level distribution in the dialectal portion of our data.

Naturally-Occurring Code-Switching. Using 1M random tweets from our data, we perform an analysis of code-switching. For this, we employ simple string matching to identify Arabic and run the CLD3 language ID tool[4] on the non-Arabic string sequences. We find the data to have 4.14% non-Arabic. These turn out to be almost always natural code-switching involving many foreign languages (e.g., English, French, Korean, etc.).

[2] https://catalog.ldc.upenn.edu/LDC2009T30.
[3] https://www.hindawi.org/books.
[4] https://github.com/google/cld3

Source	Size	Tokens
AraNews	8.6GB	847.8M
Books	650MB	72.5M
El-Khair	16GB	1.6B
Gigawords	10GB	1.1B
OSIAN	2.8GB	292.6M
OSCAR-MSA	31GB	3.4B
OSCAR-Egyptian	32MB	3.8M
Wiki	1.4GB	156.5M
MSA-Total	**70GB**	**7.1B**
Twitter (1.5B)	**178GB**	**21.9B**
ALL	**248GB**	**29.0B**

Table 1: The MSA and Twitter resources used to pre-train AraT5$_{\text{MSA}}$, AraT5$_{\text{TW}}$, and AraT5.

2.2 Pre-Processing and Vocabulary

We remove diacritics and replace URLs and user mentions with <URL> and <USER>. We also clean the data by removing HTML tags, elongation, and the hash signs. Further, we reduce repetitive characters, emojis, and emoticons to one. To create our language model vocabulary, we use SentencePiece (Kudo, 2018) to encode text as WordPiece tokens (Sennrich et al., 2016) with 110K WordPieces. To allow for further pre-training (and/or fine-tuning) on additional languages, we extract our vocabulary as follows: 70M MSA sentences, 200M Arabic twitter data, 15M sentences from Wikipedia English, and 5M sentences from the Wikipedia of 10 other languages (Bulgarian, French, German, Greek, Italian, Portuguese, Russian, Spanish, Turkish, Czech).[5] In § 3.1.2, we describe parallel data from four of these languages on which we fine-tune our models for X→Arabic MT. Our respective results (reported in Table 4.2) demonstrate the utility of including foreign vocabulary in our models.

2.3 AraT5

Model Architecture. We leverage our unlabeled MSA and Twitter data described in § 2.1 to pre-train three models: **AraT5$_{\text{MSA}}$** on MSA data, **AraT5$_{\text{TW}}$** on twitter data, and **AraT5** on both MSA and twitter data using the T5$_{\text{Base}}$ encoder-decoder architecture (Raffel et al., 2019). Each of the encoder and decoder components is similar in size and configuration to BERT$_{\text{Base}}$ (Devlin et al., 2019), with 12 layers each with 12 attention heads, and 768 hidden units. In total, this results in a model with $\sim$ 220 million parameters.[6] **Objective.** Raffel et al. (2019) pre-train T5$_{\text{Base}}$ using a self-supervised (denoising) objective. The main idea is to feed the model with masked (corrupted) versions of the original sentence, and train it to reconstruct the original sequence. Inspired by BERT's objective (Devlin et al., 2019), the denoising objective (Raffel et al., 2019) works by randomly sampling and dropping out 15% of tokens in the input sequence. All consecutive spans of dropped-out tokens are then replaced by a single sentinel token. **Pre-Training.** For all three of our pre-trained models, we use a learning rate of 0.01, a batch size of 128 sequences, and a maximum sequence length of 512, except for AraT5$_{\text{TW}}$ where the maximum sequence is 128.[7] We pre-train each model for 1M steps. Pre-training of each model took $\sim$ **80 days** on one Google Cloud TPU with 8 cores (v3.8) from TensorFlow Research Cloud (TFRC).[8] We now introduce our language generation and understating benchmarks.

3 ARGEN

In order to evaluate our pre-trained language models, we introduce our new benchmark for Arabic language generation evaluation **ARGEN**. It includes *19 different datasets* with *59 test splits* and covers *seven* tasks: machine translation (MT), code-switched translation (CST), text summarization (TS), news title generation (NGT), question generation (QG), transliteration (TR), and paraphrasing (PPH). As such, ARGEN has wide-coverage both in terms of the number of tasks and datasets. It is also linguistically diverse as it covers both MSA and various Arabic dialects, in addition to *Arabizi* (romanized Arabic in the TS task) and code-switching (in the CST task). We now describe each component of ARGEN.

3.1 Machine Translation

To design the MT component of ARGEN, **ARGEN$_{\text{MT}}$**, we consolidate 7 unique datasets with 46 different test splits. The datasets come from both MSA and Arabic dialects, and range between 600-138K sentences (details in Table C.2 in Appendix). We introduce each dataset briefly here.

3.1.1 Arabic → English

(1) United Nations Parallel Corpus. Ziemski et al. (2016) introduce this parallel corpus of man-

[5] The MSA and twitter data are extracted from our training data presented in Section 2.1.

[6] The output dimensionality is $d_{ff} = 3,072$ and inner dimensionality of $d_{kv} = 64$.

[7] We choose the same maximum sequence used in MARBERT (Abdul-Mageed et al., 2021), the most powerful model trained on Arabic twitter to date (Farha and Magdy, 2021).

[8] https://www.tensorflow.org/tfrc.

ually translated UN documents covering the six official UN languages (i.e., Arabic, Chinese, English, French, Russian, and Spanish). The corpus consists of development and test sets only, each of which comprise $4,000$ sentences that are one-to-one alignments across all official languages.

(2) IWSLT Corpus. Several Arabic-to-English parallel datasets were released during IWSLT evaluation campaigns (Federico et al., 2012; Cettolo et al., 2013, 2014, 2016). The datasets are mainly extracted from transcriptions of TED talks between 2010 and 2016, and the QCRI Educational Domain Corpus (QED 2016) (Abdelali et al., 2014).

AraBench Datasets. Sajjad et al. (2020) introduce AraBench, an evaluation suite for MSA and dialectal Arabic to English MT consisting of *five* publicly available datasets: **(3) ADPT:** Arabic-Dialect/English Parallel Text (Zbib et al., 2012), **(4) MADAR:** Multi-Arabic Dialect Applications and Resources dataset (Bouamor et al., 2018), **(5) QAraC:** Qatari-English speech corpus (Elmahdy et al., 2014), and **(6) Bible:** The English Bible translated into MSA, Moroccan, and Tunisian Arabic dialects.[9] For all these datasets, we use the same splits as Sajjad et al. (2020) in our experiments.

3.1.2 X → Arabic

To investigate ability of our models to generate Arabic starting from foreign languages in our vocabulary, we create an X→Arabic benchmark of four languages (English, French, German, and Russian) by extracting parallel data from OPUS (Tiedemann, 2012). For each language, we pick 1M sentences for training and 5K sentences for each of development and test splits. This gives us our seventh $\textbf{ARGEN}_{\textbf{MT}}$ dataset, which we call **(7) OPUS-X-Ara**.

3.2 Code-Switched Translation

There is rising interest in translating *code-switched* data (Nagoudi et al., 2021). Our purpose here is to translate Arabic text involving code-switching from a foreign language into **(i)** that foreign language as well as into **(ii)** MSA. Hence we create $\textbf{ARGEN}_{\textbf{CST}}$, our code-switched translation benchmark component, using *four* sub-test sets. Two of these are *natural* and two are *synthetic*, as follows:

Natural Code-Switched Data. We create two human written (natural) code-switched parallel datasets: **(1) ALG-CST.** This is collected from Algerian Twitter and consists of code-switched Arabic-French posts. We translate these manually into monolingual French. **(2) JOR-CST.** This is collected from Jordanian Twitter and consists of code-switched Arabic-English posts, which we manually translate into monolingual English. Each of ALG-CST and JOR-CST comprises 300 tweets (total=600). Human translation is performed by one native speaker from each dialect with semi-native English/French fluency.

Synthetic Code-Switched Data. We use the multilingual sequence-to-sequence model mBART (Liu et al., 2020) to create synthetic code-switched data following Jawahar et al. (2021). We exploit the UN multi-parallel data (Ziemski et al., 2016) using the Arabic-English and Arabic-French test splits ($4,000$ sentences each, described in § 3.1) to generate our two code-switched test sets **(3) MSA-EN** and **(4) MSA-FR**. In each case, we use mBART to translate $\sim 30\%$ random Arabic n-grams into the target language (i.e., English or French).

3.3 Text Summarization

To build our *text summarization* benchmark component, $\textbf{ARGEN}_{\textbf{TS}}$, we use the following:

Essex Arabic Summaries Corpus (EASC). EASC (El-Haj et al., 2010) contains 153 Arabic Wikipedia and newspaper articles, each with 5 human-generated extractive summaries (total=765 summaries). The summaries are crowdsourced via Mechanical Turk.[10]

WikiLingua. An abstractive summarization dataset in 18 languages, including Arabic (Faisal Ladhak and McKeown, 2020). It contains articles and their summaries from WikiHow.[11] The Arabic part includes summaries for 29.2K articles, which we split into 80% Train (23.4K), 10% Dev (2.9K), and 10% Test (2.9K).

3.4 News Title Generation

The purpose of the *news title generation (NTG)* task is to produce proper news article titles (Liang et al., 2020). We introduce NTG as a *new* task for Arabic language generation. Given an article, a title generation model needs to output a short grammatical sequence of words suited to the article content. For this, we introduce $\textbf{ARGEN}_{\textbf{NTG}}$, a novel NTG dataset exploiting 120K articles along

[9] The United Bible Societies https://www.bible.com.

[10] http://www.mturk.com/

[11] http://www.wikihow.com

with their titles extracted from AraNews (Nagoudi et al., 2020).[12] We only include titles with at least three words in this dataset. We split $\text{ARGEN}_{\text{NTG}}$ data into 80% Train (93.3K), 10% Dev (11.7K), and 10% Test (11.7K). Details about $\text{ARGEN}_{\text{NTG}}$ are in Table C.1 (Appendix). A sample of a news article from our Test split and example titles generated by our models are in Table D.5 (Appendix).

3.5 Question Generation

In the *question generation (QG)* task, a question is produced for a passage (Gehrmann et al., 2021). Given the absence of an Arabic QG dataset, we create a new Arabic QG dataset (**ARGEN_{QG}**) using a publicly available Arabic question answering (QA) resource. We follow Kriangchaivech and Wangperawong (2019) who train a model to generate simple questions relevant to passages and answers extracted from SQuAD (Rajpurkar et al., 2016). In our case, we build ARGEN_{QG} by extracting 96K (passage, answer, and question) triplets from **(1)** The Arabic QA dataset ARCD (Mozannar et al., 2019), and **(2)** three multi-lingual QA datasets: XTREME benchmark (Hu et al., 2020), MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2020), and TyDi QA (Artetxe et al., 2020).

3.6 Paraphrasing

The main goal of this task is to produce for a given Arabic sentence a *paraphrase* with the same meaning. In order to build our paraphrasing benchmark component (**$\text{ARGEN}_{\text{PPH}}$**), we use the following three datasets:

AraPara. We introduce AraPara, a new multi-domain Arabic paraphrasing dataset we create using English-Arabic parallel OPUS data (Tiedemann, 2012). AraPara covers several domains such as news, religion, politics, movies, and technology. To create a high quality machine generated paraphrase dataset, we follow four careful steps involving human validation (more details are offered in Appendix C.1). AraPara consists of 122K paraphrase pairs. We only use AraPara for model development, and hence we split it into 116K Train and 6K Dev.

Arabic SemEval Paraphrasing (ASEP). We also create a new Arabic paraphrasing dataset using three existing Arabic semantic similarity datasets released during SemEval 2017 (Cer et al., 2017). These are MSR-Paraphrase (510 pairs), MSR-Video (368 pairs), and SMTeuroparl (203 pairs). The pairs are labeled with a similarity score on a scale from 0 to 5. For our purpose, we only keep sentence pairs with a semantic similarity score ≥ 3.5 which gives us 603 pairs. We merge and shuffle all three ASEP datasets for our use.

Arabic Paraphrasing Benchmark (APB). APB is created by Alian et al. (2019). It consists of $1,010$ Arabic sentence pairs that are collected from different Arabic books. Paraphrasing was performed manually using six transformation procedures (i.e., addition, deletion, expansion, permutation, reduction, and replacement).

3.7 Transliteration.

Transliteration involves mapping a text written with orthographic symbols in a given script into another (Beesley, 1998). We use the *BOLT Egyptian Arabic SMS/Chat and Transliteration dataset* (Song et al., 2014),[13] a collection of naturally-occurring chat and short messages (SMS) from Egyptian native speakers. The messages (sources) were natively written in either romanized Arabizi or Egyptian Arabic orthography. The target is the Egyptian transliteration of these message.[14] For experiments, we use the same split proposed by Shazal et al. (2020) (58.9K for Train and 5.4K for Dev and Test each). We refer to this dataset as **ARGEN_{TR}**.

4 Evaluation on ARGEN

Baselines and Procedure. For all tasks, we compare our models to models fine-tuned with mT5 using the same training data. In addition, for MT, we compare to a vanilla sequence-to-sequence (S2S) Transformer (Vaswani et al., 2017) trained from scratch as implemented in Fairseq (Ott et al., 2019). For all models and baselines, across all tasks, we identify the best model on the respective Dev data and blind-test it on Test data. As a rule, we report on both Dev and Test sets. All our Dev results are in Section C.2 in the Appendix.

4.1 Machine Translation.

We train two S2S Transformers models on 2M (S2S_{2M}) and 10M (S2S_{10M}) MSA-English parallel sentences extracted from OPUS. We take these

[12] We ensure no overlap exists between ARGEN_{TG} and the AraNews data we use to pre-train our language models (described in § 2.3).

[13] https://catalog.ldc.upenn.edu/LDC2017T07

[14] Some transliteration sequences involve code mixing between Egyptian Arabic and English.

	Dataset	Test Split	S2S$_{2M}$	S2S$_{10M}$	mT5	AraT5$_{Tw}$	AraT5$_{MSA}$	AraT5	SOTA
DIA	ADPT†	Lev	4.30	6.20	8.33	8.32	**8.52**	8.42	10.80
		Egy	5.21	8.9	12.57	11.25	12.38	**12.92**	14.00
	Bible I	Tun.	4.12	4.44	8.08	5.86	**8.52**	7.94	7.00
		Mor.	2.60	2.80	7.21	4.69	**7.83**	6.82	4.20
	MADAR I†	Egy.	17.25	17.71	24.44	21.75	**24.98**	24.66	28.90
		Qat.	15.98	17.92	23.72	22.23	**24.00**	23.92	27.60
		Leb.	12.15	10.14	14.61	12.25	**14.92**	14.18	17.00
		Tun.	8.49	8.57	10.12	9.09	**10.18**	9.60	11.40
		Mor.	11.07	11.83	16.61	12.37	**16.99**	16.82	14.70
	MADAR II†	Egy-Alex.	19.01	19.74	29.34	24.79	**29.87**	29.02	28.90
		Egy-Asw.	16.37	16.95	23.01	19.52	**23.41**	22.06	26.30
		Sud-Kha.	24.97	25.65	30.87	28.13	**31.39**	30.65	36.70
		Yem-San.	19.62	20.35	24.87	23.19	**26.10**	25.73	29.90
		Oma-Mus.	29.12	30.66	33.74	32.15	**34.62**	34.18	39.50
		KSA-Riy.	26.14	26.66	33.54	30.81	**33.86**	33.59	40.70
		KSA-Jed.	16.08	17.21	**23.57**	20.91	23.45	23.11	27.40
		Iraq-Bag.	15.98	19.09	22.92	20.84	23.24	**22.52**	28.30
		Iraq-Bas.	16.46	17.12	**22.94**	20.47	22.61	22.00	27.70
		Iraq-Mos.	18.25	19.14	23.69	21.95	**24.41**	23.12	30.00
		Pal-Jer.	15.18	16.06	24.61	20.91	**24.95**	24.45	27.00
		Jor-Amm.	18.68	18.86	26.45	22.92	**26.78**	25.26	30.00
		Jor-Salt.	17.14	17.78	26.04	23.05	**26.56**	26.05	29.60
		Syr-Dam.	13.63	14.83	21.93	18.55	**22.54**	21.80	25.90
		Syr-Alep.	14.16	15.27	22.39	19.55	22.91	**23.26**	26.40
		Alg-Alg.	13.94	14.24	16.97	14.26	**17.46**	16.62	17.30
		Lyb-Trip.	14.49	15.44	20.17	17.56	**20.31**	19.85	22.80
		Lyb-Beng.	19.02	19.32	25.50	23.39	25.46	**25.54**	28.40
		Tun-Saf	7.89	8.57	9.26	8.15	**9.94**	9.60	10.80
		Mor-Fes	15.09	15.59	22.81	17.33	**23.33**	21.97	20.90
	QAraC†	Qatar	10.33	10.47	**11.84**	11.11	11.42	10.57	11.90
	Average DIA		14.75	15.58	20.66	18.28	**21.02**	20.49	23.49
MSA	Bible II†	Test 1	10.44	10.86	15.58	13.04	**16.38**	15.71	17.00
		Test 2	5.55	6.20	12.14	9.27	**12.53**	11.64	12.80
	MADAR I†	MSA	10.33	10.47	**11.84**	11.11	11.42	10.57	11.90
	IWSLT‡	TED10	24.12	25.13	28.02	27.35	**28.64**	28.32	28.00
		TED11	23.96	25.01	28.89	28.03	**29.93**	27.34	32.80
		TED12	28.34	28.98	33.77	32.74	**35.07**	34.238	36.50
		TED13	24.19	25.02	27.12	27.52	**27.95**	27.52	37.40
		TED14	25.64	26.48	29.85	28.64	**30.94**	30.06	31.70
		TED15	27.68	28.73	29.39	28.2	30.37	**30.45**	34.10
		TED16	25.71	25.77	28.39	27.03	**29.37**	29.18	31.80
		QED16	19.44	19.90	**21.09**	18.55	20.98	19.11	28.10
	UN††	AR-EN	52.54	53.12	52.38	51.48	53.29	**52.96**	56.90
	Average MSA		23.54	24.19	27.03	25.43	**27.77**	26.98	30.63
	Average All		19.14	19.89	23.84	21.85	**24.39**	23.74	27.06

Table 2: English to Arabic results in BLEU using ARGEN$_{MT}$ datasets. **Baseline I :** Sequence-to-Sequence Transformer models trained from scratch on 2M and 10M parallel sentences. **Baseline II :** mT5 (Xue et al., 2020). **Our models :** ArT5$_{Tweet}$, ArT5$_{MSA}$, ArT5. **SOTA :** † Sajjad et al. (2020) trained on $\sim$ 42M sentences, ‡ Durrani et al. (2017) trained on $\sim$ 59M sentences, †† Junczys-Dowmunt et al. (2016) trained on $\sim$ 12M sentences.

two models as our **baseline I**. We also fine-tune our three models as well as mT5 on the same OPUS 2M MSA-English parallel sentences used for baseline I. Fine-tuned mT5 is our second baseline **baseline II**.

Arabic → English. Results of ARGEN_{MT} are reported in Table 2. Results show that our models achieve best BLEU score in 37 out of the 42 tests splits. $\text{AraT5}_{\text{MSA}}$ acquires best results in 32 of these test splits, outperforming all the baselines (S2S_{2M}), (S2S_{10M}), and mT5 with +5.25, +4.99, and +0.45 BLEU points. These results are striking since our language models are pre-trained on Arabic data only (although they include English vocabulary and marginal amounts of code-switching; see § 2.1). In other words, even under this arguably *zero-shot* setting,[15] the models perform very well. In addition, our AraT5 model outperforms even the S2S model trained with 5X more data. For completeness, we also provide the current SOTA on each of our datasets. We do not compare our results to SOTA since these are acquired by models fine-tuned on much larger datasets than ours. For example, Sajjad et al. (2020) exploit ∼ 42M parralel sentences to train their models. To limit GPU needs during our experiments, especially given the time-consuming fine-tuning process typical of T5 models, we do not fine-tune the models on the full amounts of available parallel data. However, in the future we plan to compare our models under the full data setting.

X → Arabic. Our language models are not pre-trained on foreign data, but we include vocabulary from 11 foreign languages. Our X → Arabic experiments here are hence zero-shot (from the perspective of pre-training). Table 4.2 shows the results of $\text{AraT5}_{\text{MSA}}$ and mT5 on OPUS-X-Ara.[16] We observe that our model outperforms mT5 in the four X → Arabic sub-tasks with an average of +1.12 and +0.86 BLEU points on Dev and Test, respectively.

4.2 Code-Switched Translation.

For this task, we test on the two natural code-switched translation (CST) test sets that we manually created, ALG-FR→FR and JOR-EN→EN. We also evaluate on our two synthetic CST datasets, MSA-EN and MSA-FR, one time with EN/FR as target (e.g., MSA-EN→EN) and another with MSA as target (e.g., MSA-EN→MSA). We fine-tune our three pre-trained models as well as mT5 on the OPUS-X-Ara segments involving English and French (each with 1M parallel sentences, described in § 3.1.2), in both directions. Since these MT models are only fine-tuned on parallel monolingual data, we refer to these experiments as *zero-shot*. We test these models on both our natural and synthetic code-switched data (described in § 3.2). We report results in Table 3. Our models achieve best results in one out of the two natural test sets (with +4.36 BLEU points on ALG-FR) and *all four* synthetic test sets (e.g., +4.55 BLEU points on MSA-EN→MSA). *These results clearly show our models' remarkable language generation ability especially in the Arabic direction.*

Dataset	Split	mT5	AraT5_{Tw}	$\text{AraT5}_{\text{MSA}}$	AraT5
Natural	ALG-FR → FR	23.83	**28.19**	26.27	26.17
	JOR-EN → EN	**23.06**	21.60	21.58	20.45
Synthetic	MSA-FR → FR	12.76	10.57	**13.78**	13.25
	MSA-EN → EN	11.06	8.99	**11.53**	11.42
	MSA-FR → MSA	12.93	12.14	**14.39**	13.92
	MSA-EN → MSA	19.82	18.43	23.89	**24.37**

Table 3: Performance of our models on $\textbf{ARGEN}_{\textbf{CS}}$.

4.3 Text Summarization

For the two ARGEN_{ST} datasets, we fine-tune and identify the best model on the Train and Dev splits of WikiLingua (Faisal Ladhak and McKeown, 2020) and test on all EASC and the Test of WikiLingua. We report different ROUGE scores (Lin, 2004) in Table 5. As the Table shows, AraT5_{Tw} acquires best results on WikiLingua data, while mT5 outperforms us on EASC (we hypothesize since EASC is older data that is likely part of the mC4 on which mT5 was pre-trained). *On both datasets, we establish new SOTA* (both with our pre-trained models and mT5).

4.4 News Title and Question Generation

For both tasks, we fine-tune all our models on the Train splits of $\text{ARGEN}_{\text{NTG}}$ and ARGEN_{QG}, respectively. As Table 6 shows, *all* our models outperform mT5 on each of the two tasks. $\text{AraT5}_{\text{MSA}}$ excels with 20.61% BLEU on $\text{ARGEN}_{\text{NTG}}$ and AraT5 is at 16.99% on ARGEN_{QG}.

4.5 Paraphrasing and Transliteration

For the *paraphrasing* task, we fine-tune and validate on our new AraPra dataset and blind-test on both APB and ASEP datasets (described in§ 3.6).

[15] At best, this can be viewed as *few-shot* pre-training.

[16] To limit GPU time, we fine-tune only $\text{AraT5}_{\text{MSA}}$ model on the X→Arabic direction since it performed best on Arabic→English section above.

Dataset	DEV		TEST	
	mT5	AraT5$_{MSA}$	mT5	AraT5$_{MSA}$
EN → AR	13.60	**15.72**	17.80	**18.58**
DE → AR	12.88	**13.74**	11.92	**12.80**
FR → AR	17.52	**17.96**	18.61	**18.99**
RU → AR	26.78	**27.87**	26.63	**28.01**
Average	17.70	**18.82**	18.74	**19.60**

Table 4: Performance of MT models on OPUS-X-Ara.

Dataset	Metric	mT5	AraT5$_{Tw}$	AraT5$_{MSA}$	AraT5
EASC	Rouge1	**62.98**	60.74	59.54	54.61
	Rouge2	**51.93**	48.89	47.37	43.58
	RougeL	**62.98**	60.73	59.55	54.55
WikiLin.	Rouge1	71.63	**74.61**	72.64	73.48
	Rouge2	63.60	**67.00**	64.21	65.09
	RougeL	71.56	**74.52**	72.57	73.37

Table 5: Performance of summarization models on Test. We consider mT5 as SOTA for WikiLin, and Alami et al. (2021) (ROUGE1=59.17) for EASC.

As Table 6 shows, AraT5$_{MSA}$ is best on APB (17.52 BLEU) and ASEP (19.38 BLEU). For *transliteration*, we fine-tune our models on the Train split of ARGEN$_{TR}$. As Table 6 shows, each of AraT5$_{MSA}$ and AraT5 outperform mT5. Notably, AraT5$_{MSA}$ is at 65.88 BLEU, outperforming previous SOTA (Shazal et al., 2020) by 7.1 points.

Dataset	mT5	AraT5$_{Tw}$	AraT5$_{MSA}$	AraT5
ARGEN$_{NTG}$	19.49	20.00	**20.61**	20.51
ARGEN$_{QG}$	15.29	12.06	14.18	**16.99**
ARGEN$_{TR}$	60.81	59.55	**65.88**	62.51
ARGEN$_{PPH}$ I	19.32	18.17	**19.38**	19.03
ARGEN$_{PPH}$ II	19.25	17.34	**19.43**	18.42

Table 6: Performance of our models on title, question generation, transliteration, and paraphrasing tasks in BLEU. **ARGEN$_{PPH}$ I and II**: results on ASEP and APB paraphrase datasets, respectively. We consider mT5 as SOTA for NTG, QG, and PPH **ARGEN$_{NTG}$**, **ARGEN$_{QG}$**, and **ARGEN$_{PPH}$**. For **ARGEN$_{TR}$**, SOTA is Shazal et al. (2020) (BLEU=65.88).

4.6 Evaluation on Arabic NLU

We also evaluate our new pre-trained models on the recently proposed Arabic language understanding and evaluation benchmark, ARLUE (Abdul-Mageed et al., 2021) that involves six cluster tasks (i.e., sentiment analysis, social meaning, topic classification, dialect identification, named entity recognition, and question answering). Our models establish new SOTA on the benchmark with an *ARLUE score of* 77.52 vs. the previous SOTA of 76.53, reported by ARLUE authors. We provide results of this set of experiments in Appendix B.

5 Analysis and Discussion

5.1 Multilingual vs. Dedicated Models.

Our results confirm the utility of dedicated language models as compared to multilingual models such as mT5 (101+ languages). Our AraT5 model outperforms mT5, even though it is pre-trained with 49% less data (see § 2.1). One reason might be that massively multilingual models are more prone to suffering from capacity issues. Data quality is another challenge for multilingual models. As pointed out earlier, Kreutzer et al. (2021) find systematic issues with data representing several languages (including Arabic) in the mC4 dataset on which mT5 is pre-trained. We perform a data quality study confirming the findings of Kreutzer et al. (2021). We also find Arabic mC4 data to be less geographically diverse than our Twitter pre-training data (described in § 2.1). Our mC4 data study is in Appendix A.

Code-Switching. We also study code-switching in both our Twitter dataset and the Arabic part of mC4. We find that while our Twitter data involves natural code-switching ($\sim$ 4% of sequences), code-switching in Arabic mC4 is very rare. This explains the strong performance of our AraT5$_{Tw}$ model on the natural code-switched translation data on French. We conjecture that mT5 good performance on English code-switched data is due to it being pre-trained on very large amounts of English rather than natural code-switching.

5.2 Effect of Sample Length on MT.

We were inquisitive how MT models fine-tuning our pre-trained language models compare to mT5 under different length conditions. For this, we **(1)** merge all MSA and dialectal Test datasets in our Arabic→English experiments to form a single dataset that we then **(2)** split into three bins/Test sets based on sentence length as shown in Table D.1. As the Table shows, our AraT5$_{MSA}$ outperform mT5 in *all* but one condition (where our model acquires marginally less performance). We also performed similar evaluation on the merged Dev sets of all MSA and dialectal Arabic MT datasets in the Arabic→English direction. We do not show related results here, but we note our AraT5$_{MSA}$ outperforms mT5 on *all* conditions.

(1) Source:	J'aime une vidéo Episode 1 - نسيبتي العزيزة 4 **:ALG-FR**
Target:	**FR :** J' aime une vidéo Episode 1 - ma chère belle-mère 4
mT5	J' aime une v- Chère nièce 4.
AraT5$_{Tw}$	J'aime une vidéo Episode 1 - ma chère tante 4.
AraT5$_{MSA}$	J'aime une vidéo 1 - Ma chère **sœur** 4.
AraT5	J'aime une vidéo 1 - Ma chère **bébé**
(2) Source:	بطلة العالم في ال comfort zone وهاد شيء بائس حقيقة **:JOR-EN**
Target:	**EN :** The world champion in the comfort zone and this is really miserable
mT5	the world world champion in comfort zone, and that's really a **bad** thing.
AraT5$_{Tw}$	the world **hero** in comfort zone and it's really a **miserable** thing.
AraT5$_{MSA}$	world champion in comfort zone, and that's really a **bad** thing.
AraT5	the world's the world's **hero** in the comfort zone, and it's a really **bad** thing.

Table 7: CS sentences with their English/French translations using our Models and mT5. Data samples are extracted from the Dev datasets. **Green** refers to good translation. **Red** refers to problematic translation.

5.3 Qualitative Analysis.

We also perform qualitative analyses of the outputs of several of our models, including as to length of MT source data (Appendix D). In particular, our analyses are for the following tasks: machine translation, code-switched translation, paraphrasing, transliteration, and news title generation. **MT Model.** Table D.2 (Appendix) shows three examples of Arabic→English MT models. Sentence (1) is in **MSA source**, sentence (2) is in Levantine Arabic source, and sentence (3) is in Egyptian source. In all three examples, one or more of our models generate(s) more fluent translations than mT5. This includes ability of our models to translate dialectal sentences where mT5 seems to struggle (e.g., mT5 is not able to translate the equivalents of "drive" from Egyptian Arabic).

Code-Switched Translation Model. Table 7 shows two code-switched examples from ARGEN$_{CS}$. Sentence (1) is Algerian dialect at source translated into French, while sentence (2) is Jordanian dialect translated into English. In both cases, our models not only handle the dialects but also their use in code-switched contexts better than mT5.

Paraphrasing, Transliteration, and Title Generation. Each of Tables D.3, D.4, and D.5 (Appendix D) shows two output samples from our paraphrasing, transliteration, and title generation models, respectively. In each case, the samples are high-quality, informative, and fluent. Our paraphrase samples also tightly capture the meaning of the source sentences.

6 Related Work

Multilingual LMs. ***mBERT*** is the multilingual version of BERT (Devlin et al., 2019), which is an encoder model with bidirectional representations from Transformers trained with a denoising objective. mBERT is trained on Wikipedia for 104 languages, including Arabic. ***XLM-R*** (Conneau et al., 2020) is also a Transformer-based multilingual masked language model pre-trained on more than 2TB of CommonCrawl (CC) data in 100 languages, including Arabic (2.9B tokens). XLM-R model uses the same masking objective as BERT, but not the next sentence prediction. ***mT5*** (Xue et al., 2020) is the multilingual version of **T**ext-**t**o-**T**ext **T**ransfer **T**ransformer model (T5) (Raffel et al., 2019). T5 is an encoder-decoder Transformer similar in configuration and size to a BERT$_{Base}$. It is trained on mC4, which is $\sim$ 26.76TB for 101 languages generated from 71 CC dumps.

Arabic LMs. ***AraBERT*** (Antoun et al., 2020) is an Arabic pre-trained language model based on the BERT$_{Base}$ architecture with 24GB of MSA data. ***ARBERT*** and ***MARBERT*** (Abdul-Mageed et al., 2021) are two BERT-based models, with the first focused on MSA (61GB) and the second on both MSA and dialects (128GB). MARBERT achieves SOTA on most Arabic NLU tasks. ***QARiB*** (Abdelali et al., 2021) is similarly a BERT-based model covering both MSA and dialects. ***CamelBERT*** (Inoue et al., 2021) is also a BERT-based model pre-trained with MSA, dialectal, and classical Arabic.

7 Conclusion

We introduced three powerful Arabic-specific text-to-text Transformer models trained on large MSA and/or Arabic dialectal data. We also introduced ARGEN, a unified benchmark for Arabic Natural Language *generation* evaluation composed of *seven* tasks collected from a total of 19 datasets. Our models outperform mT5 on *all* ARGEN tasks (52 out of 59 test sets, i.e., 88.14%). This is true even for MT involving four foreign languages from which the models have seen marginal or no pre-training data (i.e., zero- and few-shot pre-training). Our models also set new SOTA on the large Arabic language *understanding* evaluation benchmark ARLUE. Our models involve vocabulary from 11 languages other than Arabic, and hence can easily be further pre-trained/fine-tuned in these languages. Our models are publicly available, and ARGEN datasets are accessible from our repository.

Acknowledgements

We gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004), Canadian Foundation for Innovation (CFI; 37771), Compute Canada (CC),[17], UBC ARC-Sockeye,[18] and Advanced Micro Devices, Inc. (AMD). We thank the Google TFRC program for providing us with free TPU access.[19] Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSERC, SSHRC, CFI, CC, ARC-Sockeye, AMD, or Google. We thank Bashar Talafha for help with code-switching data preparation.

Ethics Statement

Energy efficiency. Our models, similar to many deep learning language models, take significant pre-training time and are not energy efficient. We acknowledge this important issue and believe work on creating energy efficient models should receive scholarly attention.

Data. Our pre-training datasets are collected from the public domain and cover diverse communities. As we have demonstrated, our resulting models are better equipped to power applications involving several varieties of Arabic as well as code-switched language use involving Arabic. From this perspective, we hope they add to ongoing efforts in the community to design models that are fairer and more representative.

ARGEN Benchmark Release. We design ARGEN using both existing datasets and new datasets that we create for this work. In our accompanying GitHub repository, we link to all existing publicly available components of the benchmark with standard splits from source as well as components that can be acquired from data organizations. In addition, we released all the new datasets we have developed. While we have prioritized standardizing evaluation on as many unified and consolidated datasets and tasks as possible, we also report performance on individual test sets so as to enable the community to replicate our work even on particular parts or tasks of ARGEN if they so wish.

AraT5 Models Release. All our pre-trained models are publicly available for non-malicious use. We acknowledge our models may still be misused in real world. However, we hope the models will be deployed in domains such as education, disaster management, health, recreation, travel, etc. in socially beneficial ways. These meaningful potential use cases are behind our decision to release the models.

[17] https://www.computecanada.ca

[18] https://arc.ubc.ca/ubc-arc-sockeye

[19] https://sites.research.google/trc/about/

References

Mourad Abbas, Kamel Smaïli, and Daoud Berkani. 2011. Evaluation of topic identification methods on arabic corpora. *JDIM*, 9(5):185–192.

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The amara corpus: Building parallel language resources for the educational domain. In *LREC*, volume 14, pages 1044–1054.

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic Dialect Identification in the Wild. *arXiv preprint arXiv:2005.06557*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the ACL-IJCNLP 2021 Main Conference*. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*.

Muhammad Abdul-Mageed, Chiyu Zhang, Abdel-Rahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Nabil Alami, Mohammed Meknassi, Noureddine Ennahnahi, Yassine El Adlouni, and Ouafae Ammor. 2021. Unsupervised neural networks for automatic arabic text summarization using document clustering and topic modeling. *Expert Systems with Applications*, 172:114652.

Marwah Alian, Arafat Awajan, Ahmad Al-Hasan, and Raeda Akuzhia. 2019. Towards building arabic paraphrasing benchmark. In *Proceedings of the Sec-*

ond International conference on Data Science E-learning and Information Systems (DATA' 2019), pages 1–5.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Kenneth Beesley. 1998. Romanization, transcription and transliteration. *Retrieved June*, 19:2006.

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)*.

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 (WANLP19), Florence, Italy*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The iwslt 2016 evaluation campaign. In *International Workshop on Spoken Language Translation*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th iwslt evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation, Heidelberg, Germany*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, volume 57.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. Ant corpus: an arabic news text collection for textual classification. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142. IEEE.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and Stephan Vogel. 2017. Qcri machine translation systems for iwslt 16. *arXiv preprint arXiv:1701.03924*.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2010. Using mechanical turk to create a corpus of arabic summaries.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*.

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a TV broadcasts speech recognition system for qatari Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3057–3061, Reykjavik, Iceland. European Language Resources Association (ELRA).

Claire Cardie Faisal Ladhak, Esin Durmus and Kathleen McKeown. 2020. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Findings of EMNLP, 2020*.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Paul Michael, and Stüker Sebastian. 2012. Overview of the iwslt 2012 evaluation campaign. In *IWSLT-International Workshop on Spoken Language Translation*, pages 12–33.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi

Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2021. Exploring text-to-text transformers for english to hinglish machine translation with synthetic code-mixing. *NAACL 2021*, page 36.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. *arXiv preprint arXiv:1610.01108*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Kettip Kriangchaivech and Artit Wangperawong. 2019. Question generation by transformers. *arXiv preprint arXiv:1909.05017*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Michael McCandless. 2010. Accuracy and performance of google's compact language detector. *Blog post*.

Hussein Mozannar, Karl El Hajal, Elie Maamary, and Hazem Hajj. 2019. Neural arabic question answering. *arXiv preprint arXiv:1906.05394*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. Investigating code-mixed Modern Standard Arabic-Egyptian to English machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, Muhammad Abdul-Mageed, Tariq Alhindi, and Hasan Cavusoglu. 2020. Machine generation and detection of arabic manipulated and fake news. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 69–84.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Motaz K Saad and Wesam M Ashour. 2010. Osac: Open source arabic corpora. *Osac: Open source arabic corpora*, 10.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 167–177.

Zhiyi Song, Stephanie M Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, et al. 2014. Collecting natural sms and chat conversations in multiple languages: The bolt phase 2 corpus. In *LREC*, pages 1699–1704. Citeseer.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

Appendices

A A Study of Arabic mC4 Data Quality

Xue et al. (2020) train mT5 on the mC4 dataset. They report 57B Arabic tokens (almost double our token size) from 53M webpages, making 1.66% of all mT5 data. For our analysis, we randomly sample 1M paragraphs from the Arabic part of mC4. We use paragraphs rather than whole documents for a more fine-grained analysis that is more comparable to our own data (especially in the case of Twitter). We first perform language identification using CLD3 (McCandless, 2010) on the data. We find a sizable amount of the data (i.e., 13.59%) to be non-Arabic (mostly English or French). We manually inspect $\sim$ 100 random samples of the data predicted as non-Arabic. We find these are mostly either non-linguistic content (e.g., java-script or HTML code) or non-Arabic text. The non-Arabic text is sometimes foreign language advertising or even full translation of the Arabic text in some cases. In many cases, non-Arabic is also boilerplate text such as that in web fora. Also, no samples of the non-Arabic included real **code-switching**.

We also run an in-house MSA-dialect classifier on the same 1M data sample. The classifier predicts an overriding majority of the data (99.83%) as MSA. We again manually inspect $\sim$ 100 samples from the small fraction predicted as dialects (i.e., 0.17%). While we find some of these to be actual dialectal text (usually short belonging to either Egyptian or Saudi dialects) from web fora, in the majority of cases the text is simply names of soap operas or advertisements. Our own pre-training data in the case of Twitter, in comparison, involve much more dialectal content (28.39% as listed in § 2.1).

B Evaluation on Arabic NLU

B.1 ARLUE Benchmark

Recently, Abdul-Mageed et al. (2021) introduced ARLUE, a natural language understanding benchmark for Arabic. ARLUE is composed of 42 publicly available datasets, making it the largest and most diverse Arabic NLP benchmark. ARLUE is arranged into the six cluster tasks of sentiment analysis (SA), social meaning (SM), topic classification (TC), dialect identification (DI), named entity recognition (NER), and question answering (QA). We methodically evaluate each cluster task, ultimately reporting a single ARLUE score following Abdul-Mageed et al. (2021). Table B.1, shows a summary of the ARLUE benchmark. We briefly describe ARLUE tasks next.

ARLUE$_{\text{Senti}}$. To construct this task cluster Abdul-Mageed et al. (2021) merged 17 MSA and DA publicly available datasets.

ARLUE$_{\text{SM}}$. ARLUE$_{\text{SM}}$ refers to eight social meaning datasets covering prediction of age, dangerous speech, emotion, gender, hate speech, irony, offensive language, and sarcasm. used in this benchmark. We will follow Abdul-Mageed et al. (2021) in not merging the social meaning datasets, but rather report performance on each individual dataset as well as average performance across all tasks as part of an overall ARLUE score.

ARLUE$_{\text{Topic}}$. This benchmark component is a concatenation [20] of three topic classification datasets: Arabic News Text (ANT) (Chouigui et al., 2017), Khaleej (Abbas et al., 2011), and OSAC (Saad and Ashour, 2010).

ARLUE$_{\text{Dia}}$. Five datasets are used for dialect classification. These are AOC Zaidan and Callison-Burch (2014), ArSarcasm$_{\text{Dia}}$ (Farha and Magdy, 2020), MADAR (sub-task 2) (Bouamor et al., 2019), NADI-2020 (Abdul-Mageed et al., 2020a), and QADI (Abdelali et al., 2020).

ARLUE$_{\text{Dia}}$ involve three categories, namely, **ARLUE$_{\text{Dia-B}}$** for MSA-dialect classification (*binary*). **ARLUE$_{\text{Dia-R}}$**, and **ARLUE$_{\text{Dia-C}}$** for the region and country level classification into four classes (*region*), and 21 classes (*country*) respectively.

ARLUE$_{\text{QA}}$. Four Arabic and multilingual QA datasets are concatenated to build ARLUE$_{\text{QA}}$: ARCD (Mozannar et al., 2019) MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2020), and TyDi QA (Artetxe et al., 2020).[21]

B.2 ARLUE Evaluation

Baselines. For comparison, we fine-tune a number of models on the same training data as our new models. These include the multilingual sequence-to-sequence model mT5 (Xue et al., 2020), and the powerful Arabic-specific BERT-based model MARBERT (Abdul-Mageed et al., 2021). We note

[20] We note that the classes were straightforwardly merged without modifying any class labels.

[21] All corresponding splits from the different QA datasets are merged.

that MARBERT achieves the SOTA [22] across the majority of 6 cluster tasks of ARLUE, with the highest ARLUE score.

Settings and Evaluation. We evaluate our models on the language understanding benchmark, ARLUE, under two settings: (i) single task learning and (ii) multi-task learning. We present results on all the task clusters included in ARLUE except for NER which is a token-level task that is not straightforward with the text-to-text set up we adopt. Table B.2 shows our evaluation results using the relevant metric for each task.

Abdul-Mageed et al. (2021) introduced **ARLUE score**, a metric used to score pre-trained language model performance on multiple datasets. ARLUE score is a simply macro-average of the different scores across all task clusters, where each task is weighted equally following (Wang et al., 2018). We compute the ARLUE score (i.e., overall macro-average) for each of our three models (i.e., $\text{AraT5}_{\text{MSA}}$, AraT5_{Tw}, and AraT5) and the baseline (mT5).

Dataset	#Datasets	Task	TRAIN	DEV	TEST
$\text{ARLUE}_{\text{Senti}}$	17	SA	190.9K	6.5K	44.2K
ARLUE_{SM}	8	SM	1.51M	162.5K	166.1K
$\text{ARLUE}_{\text{Topic}}$	5	TC	47.5K	5.9K	5.9K
$\text{ARLUE}_{\text{Dia-B}}$	2	DI	94.9K	10.8K	12.9K
$\text{ARLUE}_{\text{Dia-R}}$	2	DI	38.5K	4.5K	5.3K
$\text{ARLUE}_{\text{Dia-C}}$	3	DI	711.9K	31.5K	52.1K
$\text{ARLUE}_{\text{QA}}^{\ddagger}$	4	QA	101.6K	517	7.45K

Table B.1: ARLUE categories across the different data splits. ‡ Number of question-answer pairs (Abdul-Mageed et al., 2021).

Dataset	SOTA	mT5	$\text{AraT5}_{\text{Tweet}}$	$\text{AraT5}_{\text{MSA}}$	AraT5
$\text{ARLUE}_{\text{Senti}}^{\star}$	93.30 / 94.00	92.46 / 93.50	92.79 / 93.50	**93.44 / 94.00**	93.30 / 94.00
$\text{ARLUE}_{\text{SM}}^{\dagger}$	81.60 /76.34	80.26 / 73.59	80.41 / 75.08	**81.97 / 76.60**	81.09 / 75.99
$\text{ARLUE}_{\text{Topic}}$	90.07 / 91.54	91.92 / 93.36	90.86 / 92.08	**92.32** / 93.30	**92.32 / 93.66**
$\text{ARLUE}_{\text{Dia-B}}$	88.47 / 87.87	86.48 / 85.72	87.72 / 87.06	**88.51 / 87.90**	88.01 / 87.41
$\text{ARLUE}_{\text{Dia-R}}$	90.04 / 89.67	88.30 / 87.93	90.12 / 89.65	**91.17** / 90.80	91.13 / **90.87**
$\text{ARLUE}_{\text{Dia-C}}$	47.49 / 38.53	45.94 / 38.14	53.34 / 42.02	52.65 / 42.42	**53.64 / 43.18**
$\text{ARLUE}_{\text{QA}}^{\ddagger}$	**40.47 / 62.09**	36.92 / 56.17	30.42 / 49.57	39.47 / 60.51	39.80 / 60.93
Average	75.92 / 77.15	74.61 / 75.49	75.09 / 75.56	**77.08** / 77.93	77.04 / **78.01**
$\text{ARLUE}_{\text{Score}}$	76.53	75.05	75.33	77.50	**77.52**

Table B.2: Performance of our models on ARLUE TEST datasets (Acc / F_1). * Metric for $\text{ARLUE}_{\text{Senti}}$ is Acc/ F_1^{PN}. ‡ Metric for ARLUE_{QA} is Exact Match (EM) / F_1.† ARLUE_{SM} results is the average score across the social meaning tasks. **SOTA:** MARBERT (Abdul-Mageed et al., 2021).

Single Task. We fine-tune our three models and mT5 individually on each of the six tasks of ARLUE. We typically (i.e., in *all* our experiments) identify the best checkpoint for each model on the development set, and report its performance on both development and test data. As Table B.2 shows, our AraT5 model achieves the highest ARLUE score (77.52), followed by $\text{AraT5}_{\text{MSA}}$ (77.50) and AraT5_{TW} (75.33). We note that all our models outperform mT5 and the MARBERT (SOTA) by $\sim +2.74$ and $\sim +1$ ARLUE score points, respectively.

Dataset	S/M	mT5	AraT5_{Tw}	$\text{AraT5}_{\text{MSA}}$	AraT5
$\text{ARLUE}_{\text{Dia-B}}$	S	86.48 / 85.72	87.72 / 87.06	**88.51 / 87.90**	88.01 / 87.41
	M	86.30 / 85.54	87.77 / 87.20	87.93 / 87.36	88.02 / 87.40
$\text{ARLUE}_{\text{Dia-R}}$	S	88.30 / 87.93	90.12 / 89.65	91.17 / 90.80	91.13 / 90.87
	M	89.01 / 88.15	91.53 / 91.17	91.42 / 91.15	**91.51 / 91.24**
$\text{ARLUE}_{\text{Dia-C}}$	S	45.94 / 38.14	53.34 / 42.02	52.65 / 42.42	53.64 / 43.18
	M	45.86 / 38.12	53.42 / 40.86	53.34 / 43.03	**53.70 / 43.37**

Table B.3: Performance of our models on ARLUE Dialects Test datasets on single and multi tasks setting (Acc / F_1). We copied single tasks results from Table B.2 in this table for comparison.

Dataset	S/M	mT5	AraT5_{Tw}	$\text{AraT5}_{\text{MSA}}$	AraT5
Age	S	60.86 / 61.05	62.29 / 62.48	63.26 / 63.41	63.50 / 63.66
	M	61.37 / 61.47	63.92 / 64.10	63.84 / 38.41	**63.82 / 63.93**
Dangerous	S	81.75 / 64.52	77.68/ 63.52	82.50 / 66.93	75.41 / 62.41
	M	79.03 / 66.46	**84.92** / 68.73	84.46 / **71.62**	77.53 / 66.53
Emotion	S	72.90 / 71.34	73.65 / 72.19	74.92 / 73.30	**76.51 / 75.24**
	M	70.88 / 68.87	72.79 / 71.24	74.39 / 73.08	74.28 / 72.57
Gender	S	72.05 / 71.83	72.27 / 72.06	73.83 / 73.56	73.38 / 73.24
	M	72.72 / 72.42	74.58 / 74.39	74.33 / 74.23	**74.65 / 74.52**
Hate	S	95.70 / 78.96	96.45 / 81.75	**96.95 / 84.88**	96.55 / 83.33
	M	95.75 / 79.29	97.00 / 82.73	96.40 / 82.07	96.15 / 80.39
Irony	S	82.61 / 82.40	82.48 / 82.25	83.23 / 83.05	**82.98 / 82.80**
	M	80.99 / 80.78	82.86 / 82.65	82.86 / 82.66	82.36 / 82.21
Offensive	S	91.35 / 85.93	94.40 / 90.96	94.15 / 91.10	93.80 / 90.11
	M	90.30 / 85.15	93.70 / 90.41	**94.10** / 90.83	94.05 / **90.85**
Sarcasm	S	84.83 / 72.66	84.08 / 75.42	86.92 / 76.53	**86.59 / 77.13**
	M	84.64 / 74.06	85.55 / 75.25	86.26 / 77.06	86.26 / 76.63
ARLUE_{SM}	S	80.26 / 73.59	80.41 / 75.08	81.97 / **76.60**	81.09 / 75.99
	M	79.46 / 73.56	81.92 / 76.19	**82.08** / 73.75	81.14 / 75.95

Table B.4: Performance of our models on ARLUE social meaning (SM) Test datasets on single- and multi-tasks setting (Acc / F_1). **S**: Single Task. **M**:Multi-task.

Multitask. We also investigate multitask learning (Caruana, 1997; Ruder, 2017) with our AraT5 models. This approach consists of training the model on multiple tasks simultaneously (i.e., the model and its parameters are shared across all tasks) in order to eventually improve performance on each individual task. In our case, we fine-tune our models on many tasks at the same time using: (i) The three dialect datasets: $\text{ARLUE}_{\text{Dia-B}}$, $\text{ARLUE}_{\text{Dia-R}}$, and $\text{ARLUE}_{\text{Dia-C}}$ and (ii) the social meaning datasets

[22] MARBERT outperform both multilingual encoder-only Transformers mBERT, $\text{XLM-R}_{\text{Base}}$, $\text{XLM-R}_{\text{Large}}$, and Arabic-specific BERT-based AraBERT (Antoun et al., 2020), ARBERT (Abdul-Mageed et al., 2021).

of ARLUE_{SM}. Table B.3 and Table B.4 show the results of multi-task experiments for dialect settings and social meaning, respectively. Our results show that multi-task training outperforms single task models in the majority of the dialects experiments (n=7 out of 9 experiments, 77.78% of the tasks) and half of the social meaning tasks (n=18 out of 36 experiments, 50% of the tasks). These results are promising, and hence we plan to further investigate multi-task learning with our new models in the future.

C ARGEN

C.1 Arabic Paraphrase Data

AraPara. is a new multi-domain Arabic paraphrasing dataset we create using English-Arabic parallel OPUS data (Tiedemann, 2012). To ensure high-quality, we follow four careful steps: **(1)** We pick 1 million English-Arabic parallel sentences from OPUS (Tiedemann, 2012) covering the different domains. **(2)** We translate the English sentences using a high-quality in-house English→Arabic MT model. **(3)** We run the multi-lingual semantic similarity model from Yang et al. (2019) on the Arabic machine translated sentences and the human translation (i.e., original Arabic sentences from OPUS), keeping only sentences with an arbitrary semantic similarity score between 0.70 and 0.99. This allows us to filter out identical sentence pairs (i.e., similarity score = 1) and those that are not good translations (i.e., those with a semantic similarity score < 0.70). **(4)** In order to maximize syntactic and lexical diversity of the pairs of paraphrased sentences, we perform an analysis based on word overlap between the semantically similar pair sentences (i.e., the output of the previous step). We then perform a *manual* analysis of the data, identifying sentences with unigram token overlap between 35% and 70% as sufficiently distinct paraphrase pairs. This gives us 122K paraphrase pairs. We split these sentence pairs into 116K for training and 6K for validation.

C.2 Evaluation on DEV

In this section we describe the ARGEN_{MT} datasets splits and report the evaluation results in validation datasets. Details about $\text{ARGEN}_{\text{NTG}}$ are in Table C.1 and ARGEN_{MT} datasets splits are shown in Table C.2. Moreover, The evaluation on validation datasets for ARGEN_{TS} are described in Table C.3 and C.4, respectively. Finally, Table C.5 shows the validation results of $\text{ARGEN}_{\text{NTG}}$, ARGEN_{QG}, ARGEN_{TR}, and $\text{ARGEN}_{\text{PHP}}$ datasets.

Split	Article/Title	Avg article len	Avg title len
TRAIN	93.3K	256.46	10.06
DEV	11.7K	253.11	10.03
TEST	11.7K	260.32	10.03
Total	116.6K	256.63	10.04

Table C.1: Main characteristics of $\text{ARGEN}_{\text{NTG}}$ data splits. For each split, we provide the number of article-title pairs and the average length of the articles and titles.

D Qualitative Analysis of Models

In this section, we explore ability of our models to generate MSA and dialectal Arabic under various conditions. We now overview various types of analyses in this regard. While samples presented here are handpicked, we note that they are mostly representative of outputs from our models since we mainly chose them to demonstrate different linguistic attributes that we believed would be relevant to the analysis.

Effect of Sample Length on MT. We were inquisitive how **MT models** fine-tuning our pre-trained language models compare to mT5 under **different length conditions**. For this, we **(1)** merge all MSA and dialectal Test datasets in our Arabic→English experiments to form a single dataset that we then **(2)** split into three bins/Test sets based on sentence length as shown in Table D.1. As the Table shows, our $\text{AraT5}_{\text{MSA}}$ outperform mT5 in *all* but one condition (where our model acquires marginally less performance). We also performed similar evaluation on the merged Dev sets of all MSA and dialectal Arabic MT datasets in the Arabic→English direction. We do not show related results here, but we note our $\text{AraT5}_{\text{MSA}}$ outperforms mT5 on *all* conditions.

MT Model Output. Table D.2 shows three examples of Arabic→English MT models. Sentence (1) is in **MSA source**, sentence (2) is in Levantine Arabic source, and sentence (3) is in Egyptian source. In all three examples, on or more of our models generate(s) more fluent translations than mT5. This includes ability of our models to translate dialectal sentences where mT5 seems to struggle (e.g., mT5 is not able to translate the equivalents of “drive" from Egyptian Arabic).

Code-Switched Translation Model Output. Table 7 shows two code-switched examples from $\textbf{ARGEN}_{\textbf{CS}}$. Sentence (1) is Algerian dialect at source translated into French, while sentence (2)

Varieties	Dataset	Region	Country-Level	City-Level	DEV	TEST
DIA	ADPT Zbib et al. (2012)	Levantine	-	-	-	138K
		Nile	Egypt	-	-	38K
	Bible I	Maghrebi	Tunisia	-	-	600
			Morocco	-	-	600
	MADAR I Bouamor et al. (2018)	Nile	Egypt	Cairo	-	6.5k
			Egypt	Alexandria	-	2k
			Egypt	Aswan	-	2k
			Sudan	Khartoum	-	2k
		Gulf	Qatar	Doha	-	6.5k
			Yemen	Sana'a	-	2k
			Oman	Muscat	-	2k
			KSA	Riyadh	-	2k
			Jedd	Muscat	-	2k
			Iraq	Baghdad	-	2k
			Iraq	Basra	-	2k
			Iraq	Mosu	-	2k
		Leventian	Lebanon	Beirut	-	6.5k
			Palestine	Jerusalem	-	2k
			Jordan	Amman	-	2k
			Jordan	Salt.	-	2k
			Syria	damascus	-	2k
			Syria	Alep	-	2k
		Maghrebi	Algeria	Alger	-	2k
			Lybia	Trip	-	2k
			Lybia	Beng	-	2k
			Tunisia	Tunis	-	6.5k
			Tunisia	Safax	-	2k
			Morocco	Fes	-	6.5k
			Morocco	Rabat	-	2k
MSA	Bible II	-	-	-	-	600
		-	-	-	-	600
	MADAR II Bouamor et al. (2018)	-	-	-	-	6.5k
	IWSLT TED15 Cettolo et al. (2016)	-	-	-	-	1.1k
	IWSLT TED16 / Cettolo et al. (2016)	-	-	-	-	1.1k
	IWSLT QED16 (Cettolo et al., 2016)	-	-	-	-	550
	UN Ziemski et al. (2016)	-	-	-	4k	4k
	OPUS-X-Ara	-	-	-	5k	5k

Table C.2: Arabic to English datasets included in ARGEN_{MT}. **MADAR I:** corpus consists of 2k sentences (Test) of 21 city-level dialects each. **MADAR II:** 12k sentences (5.5k for Dev, and 6.5k for Test sets) each of five other city-level dialects and MSA. **Bible I:** 600 sentences each as Dev and Test sets for Moroccan, Tunisian, and MSA. **Bible II:** Two Dev and Test splits (600 sentences each) are used for Bible MSA.

	Dataset	Test Split	S2S$_{2M}$	S2S$_{10M}$	mT5	AraT5$_{Tw}$	AraT5$_{MSA}$	AraT5	SOTA
DA	ADPT†	Lev	4.90	7.50	10.12	**10.53**	9.33	9.53	11.00
		Egy	5.04	9.21	11.63	10.68	11.33	**11.87**	13.40
	Bible I†	Tun.	4.44	4.80	6.98	4.63	**7.48**	6.50	7.20
		Mor.	3.22	3.47	**7.65**	5.98	8.25	7.83	4.10
	MADAR I†	Egy.	17.1	17.71	24.07	21.68	**24.75**	24.29	27.1
		Qat.	16.52	17.92	23.45	22.32	**23.98**	23.58	28.10
		Leb.	9.61	12.93	18.19	16.06	**18.64**	16.82	21.80
		Tun.	9.06	9.30	10.62	9.23	**10.97**	10.25	12.10
		Mor.	8.46	8.40	11.83	8.39	**12.09**	11.26	10.00
	QAraC†	–	10.31	10.46	11.87	10.73	**11.30**	10.64	11.70
MSA	Bible II†	Test 1	11.43	11.33	15.68	13.13	**16.43**	15.89	16.60
		Test 2	5.88	6.41	12.76	9.69	**13.53**	11.96	12.9
	MADAR I†	MSA	40.75	41.84	39.11	38.06	**39.92**	39.25	45.8
	IWSLT‡	QED16	28.39	29.04	29.18	28.59	**30.19**	29.97	–
	UN††	Ar-En	51.54	51.97	50.84	50.14	**52.11**	51.54	–
		Average	14.67	15.66	18.50	16.94	**18.90**	18.31	17.06

Table C.3: ARGEN$_{MT}$ datasets on Dev splits. **S2S:** Sequence-to-sequence Transformer models trained from scratch without use of a language model. **SOTA:** †(Sajjad et al., 2020), ‡(Durrani et al., 2017), ††(Junczys-Dowmunt et al., 2016).

Dataset	Metric	mT5	AraT5$_{Tweet}$	AraT5$_{MSA}$	AraT5
WikiLin.	Rouge1	71.03	**74.20**	72.64	73.87
	Rouge2	62.87	**66.37**	64.24	65.76
	RougeL	70.99	**74.14**	72.55	73.79

Table C.4: Performance of our models on document summarization Dev splits.

Dataset	mT5	AraT5$_{Tweet}$	AraT5$_{MSA}$	AraT5
ARGEN$_{NTG}$	19.22	19.38	**20.19**	20.01
ARGEN$_{QG}$	13.95	11.25	12.96	**15.36**
ARGEN$_{TR}$	64.81	62.95	**69.30**	65.54
ARGEN$_{PHP}$	30.70	31.54	**33.15**	32.36

Table C.5: Performance of our models on title, question generation, transliteration, and paraphrasing DEV split based on Bleu score.

Jordanian dialect translated into English. In both cases, our models not only handle the dialects but also their use in code-switched contexts better than mT5.

Paraphrasing, Transliteration, and Title Generation Output. Tables D.3, D.4, and D.5 each shows two output samples from our paraphrasing, transliteration, and title generation models, respectively. In each case, the samples are high-quality, informative, and fluent. Our paraphrase samples also tightly capture the meaning of the source sentences.

Dataset	mT5	AraT5$_{Tweet}$	AraT5$_{MSA}$	AraT5
All Length				
MSA	28.38	27.03	**29.16**	28.65
DA	20.19	17.73	**20.54**	20.10
All	21.14	18.83	**21.55**	21.09
Sequence length < 10				
MSA	35.73	35.50	**36.96**	36.44
DA	20.81	18.73	**21.29**	20.68
All	21.70	19.75	**22.23**	21.65
$20 \leq$ Sequence length ≤ 10				
MSA	26.18	24.31	**26.90**	26.24
DA	19.74	16.30	**19.78**	19.56
All	21.03	17.94	**21.22**	20.91
$20 <$ Sequence length				
MSA	**19.50**	16.91	19.28	19.45
DA	13.51	11.52	**13.69**	13.44
All	15.20	13.05	**15.26**	15.13

Table D.1: Sequence length based results on ARGEN$_{MT}$ Test datasets.

(1) Source:	**MSA:** هل تعرفون أن أحد المتع الكبيرة للسفر وأحد مباهج أبحاث الإثنوجرافيا في فرصة العيش بين أولئك الذين لم ينسوا الأساليب في الرياح ويلمسونه في الأحجار التي صقلتها الأمطار ويتذوقونه في أوراق النباتات المرة
Target:	**EN:** Do you know that one of the intense pleasures of travel and one of the delights of ethnographic research is the opportunity to live amongst those who have not forgotten the old ways, who still feel their past in the wind, touch it in stones polished by rain, taste it in the bitter leaves of plants.
mT5	you know, one of the great **enjoyments** of travel and one of the pleasure ofs statistics research is the **opportunity** to live among those who have not forgotten old methods, who still feel their past in wind, touch the rain-saving stones and taste it in the snail of plants.
AraT5$_{Tw}$	you know, one of the big pleasures of travel and one of the physical research approaches is a living **chance** among those who have not forgetted old methods, who still feel their past in the wind, touch it in the stones that rained and taste it in the fresh plant leaves .
AraT5$_{MSA}$	**Do you know that** one of the great **pleasures** of travel and one of the joys of **ethnographic** research is the **opportunity** to live among those who have not forgotten the ancient methods, who still feel their past in the wind, touch it in rain-purified stones and taste it in the bitter leaves of plants ?
AraT5	you know, one of the great benefits of travel and one of the physiology research is the opportunity to live among those who have not forgotten the old methods, who still feel their past in the wind, they feel their past in the **stones that are refined by rain**, and they taste it in the leaf.
(2) Source:	**LEV:** عمفتش على مطعم رايق و حلو للشوي . بتحط عليهن إشارة عهالخريطة؟
Target:	**EN:** I'm looking for a nice, quiet grill-type restaurant. would you point them out on this map?
mT5	You find a nice and sweet cooking restaurant with a map sign?
AraT5$_{Tw}$	a snack on a nice and sweet sweat restaurant snack , you put on them a map sign?
AraT5$_{MSA}$	You're **looking** at a nice and sweet snack restaurant with a sign on the map?
AraT5	looking for a nice and sweet restaurant to eat , put a sign on them for the map?
(3) Source	**EGY:** ده فعلا مختلف ان الواحد يسوق على جنب الشارع اليمين.
Target:	**EN:** It's really different driving on the right side of the street.
mT5	that's really different that one walks on the right side of the street.
AraT5$_{Tw}$	that's really different that one **drives** by the right side of the street.
AraT5$_{MSA}$	That's really different that one **runs** on the right side of the street.
AraT5	That's really different that one **drives** on the right side of the street.

Table D.2: MSA and DIA sentences with their English translations using our Models and mT5. Data samples are extracted from the Dev datasets. **Green** refers to good translation. **Red** refers to problematic translation.

(1) Source	**Arabic sent:**	أولا ، تبسيط و توضيح هذه المعاهدات
Target	**Paraphrase:**	في البداية ، يجب تبسيط و توضيح هذه المعاهدات .
mT5		أولا، تبسيط وتوضيح المعاهدات.
AraT5$_{Tw}$		أولا ، تبسيط المعاهدات وتوضيحها.
AraT5$_{MSA}$		أولا، تبسيط و توضيح هذه المعاهدات.
AraT5		أولا ، تم تبسيط المعاهدات وتوضيحها.
(2) Source	**Arabic sent:**	قد يواجه عقوبة الإعدام إذا ما تمت إدانته بتهمة التجسس .
Target	**Paraphrase:**	يمكن لتهم التجسس ومساعدة العدو أن تودي لعقوبة الإعدام.
mT5		ويمكن أن يواجه عقوبة الإعدام إذا ادعى أن هناك جريمة تجسس.
AraT5$_{Tw}$		يمكن أن يحكم عليه بالإعدام إذا أدين بالتجسس.
AraT5$_{MSA}$		وقد يعاقب بالإعدام إذا أدين بتهمة التجسس.
AraT5		ويجوز الحكم عليه بالإعدام إذا أدين بالتجسس.

Table D.3: Paraphrasing examples extracted from Dev data splits.

(1) Source	**Arabizi:**	Tab matsha3'ali 5edmt el iphone men Vodafone
Target	**Egy:**	طيب ما تشغلى خدمة ال iphone من Vodafone
mT5		طب ما تشغلي خدمة الأيفون من Vodafone
AraT5$_{Tw}$		طيب ما تشغلي خدمة الأيفون من Vodafone
AraT5$_{MSA}$		طب ما تشغلي خدمة الأيفون من فودافون
AraT5		طب ما تشغلي خدمة الأيفون من فودافون
(2) Source	**Arabizi:**	3amla eih enty weih 2a5bar el 5otoba 3la 7eseek?
Target	**Egy:**	عاملة إيه انتي وإيه اخبار الخطوبة على حسك؟
mT5		عاملة إيه انتي فيه اخبار الخطوبة على حسابك؟
AraT5$_{Tw}$		عاملة إيه انتي وإيه اخبار الخطوبة على حسبك؟
AraT5$_{MSA}$		عاملة إيه انتي وإيه اخبار الخطوبة على حسابك؟
AraT5		عاملة إيه انتي وإيه اخبار الخطوبة على بحسبك؟

Table D.4: Transliteration examples extracted frm from Dev data splits.

(1) Document:	السودان اليوم : اصدر المجلس القومي للمناطق والاسواق الحره برئاسه دكتور مدثر عبدالغني عبدالرحمن وزير الاستثمار قرارا بإلغاء ترخيص عمل شركه قلب العالم الاقتصادية بجزيره مقرسم بولايه البحر الاحمر ووجه القرار الجهات المختصة بضرورة تنفيذه حيث اتخذ المجلس القرار في اجتماعه الذى انعقد بتاريخ ١٣ من يونيو الحالى....
Gold Title:	المجلس القومي الأسواق الحرة.. اصدر قرار بالقاء ترخيص عمل شركة قلب العالم
mT5:	قرار بإلغاء ترخيص عمل شركة قلب العالم الاقتصادية
AraT5$_{Tweet}$:	وزير الاستثمار يلغي ترخيص عمل شركة قلب العالم الاقتصادية بجزيره
AraT5$_{MSA}$:	إلغاء ترخيص شركة قلب العالم الاقتصادية
AraT5:	إلغاء ترخيص عمل شركة قلب العالم الاقتصادية
(2) Document:	قال وزير الطاقة التركي فاتح دونميز اليوم الجمعة، إن بلاده حصلت على إعفاء من نحو ٢٥ % من العقوبات النفطية التي فرضتها الولايات المتحدة على إيران، بما يعادل نحو ٣ ملايين طن من النفط سنويا. وقال دونميز في مقابلة مع محطة تلفزيون .....
Gold Title:	وزير تركي: إعفاء تركيا بنسبة ٢٥ % من العقوبات النفطية على إيران
mT5:	تركيا تعفي ٢٥ % من العقوبات النفطية على إيران
AraT5$_{Tweet}$:	تركيا تعفي من العقوبات النفطية بنسبة ٢٥ % على إيران
AraT5$_{MSA}$:	تركيا تحصل على إعفاء من ٢٥ % من العقوبات النفطية الأمريكية على إيران
AraT5:	تركيا تحصل على إعفاء ٢٥ % من العقوبات الأمريكية على إيران

Table D.5: Title generation samples from Dev set using our Models.