AraBench: Benchmarking Dialectal Arabic-English Machine Translation

Hassan Sajjad Ahmed Abdelali Nadir Durrani Fahim Dalvi
{hsajjad,abdelali,ndurrani,faimaduddin}@hbku.edu.qa
Qatar Computing Research Institute, HBKU Research Complex, Doha 5825, Qatar

Abstract

Low-resource machine translation suffers from the scarcity of training data and the unavailability of standard evaluation sets. While a number of research efforts target the former, the unavailability of evaluation benchmarks remain a major hindrance in tracking the progress in low-resource machine translation. In this paper, we introduce **AraBench**, an evaluation suite for dialectal Arabic to English machine translation. Compared to Modern Standard Arabic, Arabic dialects are challenging due to their spoken nature, non-standard orthography, and a large variation in dialectness. To this end, we pool together already available Dialectal Arabic-English resources and additionally build novel test sets. AraBench offers 4 coarse, 15 fine-grained and 25 city-level dialect categories, belonging to diverse genres, such as media, chat, religion and travel with varying level of dialectness. We report strong baselines using several training settings: fine-tuning, back-translation and data augmentation. The evaluation suite opens a wide range of research frontiers to push efforts in low-resource machine translation, particularly Arabic dialect translation. The evaluation suite[1] and the dialectal system[2] are publicly available for research purposes.

1 Introduction

Modern Standard Arabic (MSA) is the lingua franca of the Arab world. However, spoken language, aka dialects, are mainly used in daily interactions and social media, whereas MSA is prominently used in news, education and print media. Dialects differ widely from MSA in terms of lexical choice, morphology and syntax. While a lot of work has been done on MSA-to-English machine translation (MT), little effort has been put in translating Arabic dialects to English. Even the existing few attempts in translating Arabic dialects to English (Zbib et al., 2012; Sajjad et al., 2013) are limited to a small number of dialects. Moreover, the results of these studies are not comparable because of the lack of standard evaluation sets.

Training data and evaluation benchmarks are two essential ingredients to achieve a high quality translation system (Guzmán et al., 2019). Creating such resources for dialectal Arabic is difficult and poses a unique set of challenges: their spoken nature, non-standard orthography, variation in the level of dialectness and the fuzzy boundaries between various dialects makes it extremely difficult to create a dataset that covers all these diverse aspects. These issues have contributed towards the slow progress in dialectal Arabic MT.

Recently, several works have addressed the problem of limited training data for MT (Lample et al., 2018; Artetxe et al., 2019). However, the unavailability of standard evaluation sets remains a major barrier to track the progress of translation systems. Guzmán et al. (2019) highlighted the importance of evaluation benchmarks in pushing efforts on low-resource MT and prepared evaluation sets of Nepali-English and Sinhala-English. In this paper, we aim to further this cause by providing an evaluation benchmark for another low-resource language set, dialectal Arabic. To this end, we pool together already available dialectal Arabic-English resources and we additionally build novel testsets. We consolidate the collected data into an evaluation suite covering 4 broad, 15 fine-grained and 25 city-level dialect

[1] http://alt.qcri.org/resources/mt/arabench/

[2] https://mt.qcri.org/api

Proceedings of the 28th International Conference on Computational Linguistics, pages 5094–5107
Barcelona, Spain (Online), December 8-13, 2020

categories, belonging to diverse genres, such as media, chat, religious, travel, and varying dialectness level, thus broadening the scope of research in low-resource MT, particularly, Arabic dialect translation. We aim to increase the evaluation sets further in the future.

We carried out an extensive evaluation around our test suite by training dialectal systems from scratch and by adapting an industrial-scale MSA-to-English system towards dialectal Arabic. More concretely, we explored fine-tuning, back-translation, and data augmentation as our training strategies. A few notable findings include: i) fine-tuning an MSA-English system with a small amount of dialectal data achieves significantly better results; ii) mixing of available dialectal and MSA training data enables zero-shot translation of dialects with no training data; iii) a single general system can be built that translates a large variety of dialects and MSA effectively.

The contribution of our work are as follows:

- A suite of multi-faceted dialect-English evaluation sets, covering a diverse range of dialects, genre, and level of dialectness
- An approximate metric to measure the dialectness level of a corpus
- A large scale evaluation of dialectal Arabic-English machine translation, presenting one-hat-fits-all solution covering all variety of Arabic dialects and MSA

2 Data Curation

In this section, we describe the datasets that we consolidated to form an evaluation suite. We gather resources from previously published work as well as curate our own evaluation sets.

Arabic-Dialect/English Parallel Text (APT) Zbib et al. (2012)[3] provided Egyptian-English and Levantine-English parallel corpus consisting of ≈3.5 million tokens of Arabic dialects – 38k Egyptian-English and 138k Levantine-English sentences. The data was collected from dialectal Arabic weblogs and online user groups, and the translation was carried out by Arabic annotators using Amazon Mechanical Turk. Several research works (Zbib et al., 2012; Sajjad et al., 2013; Salameh et al., 2015) have used this data to build dialectal Arabic to English MT systems. However, all the mentioned work used different splits, due to which the results are not comparable. We refer to this data as *APT* hereafter.

Multi-dialectal Parallel Corpus of Arabic (MDC) Bouamor et al. (2014) selected 2k parallel Egyptian-English sentences from the corpus built by Zbib et al. (2012) and tasked native speakers of Palestinian, Syrian, Jordanian and Tunisian to translate these Egyptian sentences into their own dialect. They additionally translated 8000 more sentences from Egyptian-English data into Syrian only. The dataset serves as an interesting resource comparing identical sentences translated into different dialects.

MADAR Corpus Bouamor et al. (2018) selected 2k English sentences from the BTEC corpus, originally a Japanese/English bank for parallel phrases, and recruited native speakers from 26 Arab cities to translate English phrases into their own dialect. Additionally for five selected cities: Doha, Beirut, Cairo, Tunis, and Rabat, they translated 10k more sentences. The MADAR corpus is unique in terms of the variety of dialects covered and the travel domain. It was primarily built for Arabic dialect identification. Here we have geared it towards Machine Translation benchmarking.

QCA speech corpus (QAraC) Elmahdy et al. (2014) compiled a parallel corpus comprising of 14.7k Qatari-English sentences collected from Qatari TV series and talk-show programs. Al-Mannai et al. (2014) exploited the data to build a Qatari Arabic-English translation system.

The Bible The Bible has been translated into more than 3300 languages. Although a number of translations are available in Arabic, translations into Arabic dialects are still very scarce. We were able to obtain translations of the New Testament into Moroccan[4] and Tunisian[5] dialects and in MSA. Each data consists of about 8.2k parallel sentences.

[3] LDC2012T09 https://catalog.ldc.upenn.edu/LDC2012T09

[4] The Morocco Bible Society https://www.biblesociety.ma

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

Corpus	APT		MDC					MADAR	
Dialect	Nile	LEV	LEV	LEV	LEV	MGR	MSA	Nile	Nile
SC	eg	-	sy	jo	ps	tn	-	eg-Cairo	eg-Alex.
Sent.	1000	1000	1000	1000	1000	1000	1000	6500	2000
Tok. (ar)	9220	8333	9611	8167	8506	8746	9983	38304	11670
Tok. (en)	13052	11645	13052	13052	13052	13052	13052	49790	15262

Corpus	MADAR								
Dialect	Nile	Nile	Gulf	Gulf	Gulf	Gulf	Gulf	Gulf	Gulf
SC	eg-Aswan	sd-Khar.	qa-Doha	ye-Sana'a	om-Muscat	sa-Riyadh	sa-Jedd	iq-Bagh.	iq-Basra
Sent.	2000	2000	6500	2000	2000	2000	2000	2000	2000
Tok. (ar)	11998	11808	34422	11268	11652	11102	10581	10771	10307
Tok. (en)	15262	15262	49790	15262	15262	15262	15262	15262	15262

Corpus	MADAR								
Dialect	Gulf	LEV	LEV	LEV	LEV	LEV	LEV	MGR	MGR
SC	iq-Mosul	lb-Beir.	jo-Amm.	jo-Salt$	sy-Dama.	sy-Alep.	ps-Jeru.	dz-Algi.	ly-Trip.
Sent.	2000	6500	2000	2000	2000	2000	2000	2000	2000
Tok. (ar)	11249	35137	11745	12490	10636	10579	10996	11642	11537
Tok. (en)	15262	49790	15262	15262	15262	15262	15262	15262	15262

Corpus	MADAR						QAraC	Bible	
Dialect	MGR	MGR	MGR	MGR	MGR	MSA	Gulf	MGR	MGR
SC	ly-Beng.	tn-Tunis	tn-Sfax	ma-Rabat	ma-Fes	ms-msa	qa	ma	tn
Sent.	2000	6500	2000	6500	2000	6500	6713	600	600
Tok. (ar)	11567	35600	10639	38599	11690	51695	43842	9427	9312
Tok. (en)	15262	49790	15262	49790	15262	49790	57833	13808	13808

Corpus	Bible		Media				
Dialect	MSA	MSA	MGR	LEV	Gulf	MSA	MSA
SC	-	-	ma	lb	om	-	-
Sent.	600	600	526	250	467	637	621
Tok. (ar)	9061	7545	8282	3901	6520	8932	9282
Tok. (en)	13808	13808	8802	4952	7976	11817	11107

Table 1: Testsets statistics: Corpus mentions the original dataset of each set, Dialect is the coarse dialect category and Subclass (SC) is the fine-grained classification of dialect. The MADAR set additionally have city-level categorization as appended in the Subclass category (see Appendix A.1 for the complete table)

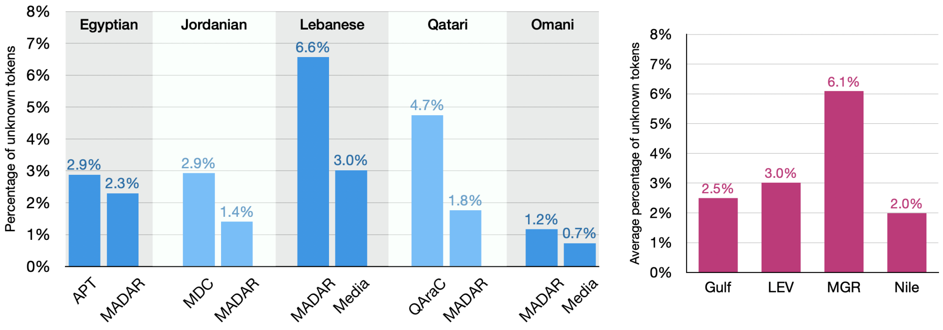
Media Testsets We prepared our own test sets for the purpose of evaluating the dialectal Arabic automatic speech recognition and MT systems, under a multilingual media monitoring project. We recorded five public broadcasting channels for two days resulting in 48 hours of recordings for each channel. Random six 15-minute long segments were selected per channel to build the test sets. This amounts to 7.5 hours of recordings that cover programs with Maghrebi, Lebanese, Omani dialects and MSA with genre involving movies, news reports and cultural programs. The recordings were transcribed and translated by a professional translation house. The testsets bring a good representation of spoken form of Arabic dialects involving a variety of genres. We refer to this collection as *Media* testsets later on.

3 Data Preparation

Data Processing We ran the following preprocessing pipeline: filtered out sentences with length above 100 words, removed diacritization and normalized Arabic characters such as replacing different forms of Alif and Hamzah (أ O, إ I, آ M) are replaced with a bare Alif ا A. Similarly for ى Y to ي y, and ة p to ه h, and converted Indian digits ٩٨٧٦٥٤٣٢١٠ to Arabic numerals 0-9 (Bouamor et al., 2014).

3.1 Data Splits

We randomly split the **APT**-lev data into 1k development set and 1k test set and used the remaining for training. **MDC** is initially selected from APT-eg (Egyptian-English corpus) of Zbib et al. (2012). In order to have consistent splits and to avoid any overlap between the evaluation sets and the training sets of any dialect, we choose the same split of 2k sentences selected by Bouamor et al. (2014) from APT-eg. We divide the selected 2k from APT-eg and the Jordanian, Syrian, Palestinian, Tunisian, MSA part of MDC



(a) Comparing DL between dialects coming from different data sources

(b) DL per coarse dialect classes

Figure 1: Dialectness level (DL)

into identical 1k splits of development and test sets. MDC additionally contains 8k Syrian sentences. We reserve them for training.

The **MADAR** corpus consists of 2k sentences of 21 city-level dialects each, and 12k sentences each of 4 other city-level dialects and of MSA. The corpus consists of small sentences/phrases with an average sentence length of 7 words. We reserve the full MADAR corpus for evaluation purposes. For the five cities (including MSA) with 12k sentences, we split them in to 5.5k sentences for development, and 6.5k sentences for test sets. For the cities with only 2k sentences, we select them for testing only. We then map each city to their respective country provided in ISO 3166-1 alpha-2 code.[6] We further combine the country-level data into four major dialect categories – Nile, Levantine, Gulf and Maghrebi.

The **QAraC** corpus consists of 14.7k parallel sentences. After cleaning, the corpus is reduced to 13.4k sentences. We split it equally into a development set and a test set.

The **Bible** corpus is a multi-parallel corpus between English, Moroccan, Tunisian, and MSA. To preserve the multi-parallelism, we choose identical random splits of 600 sentences each as a development set and a test set for each dialect. The remaining 7,019 sentences are kept for training. Lastly, the **Media** test sets come in five channel-wise categories. Since the data is of small size, we limit it for test purposes only.

3.2 Final Evaluation Sets

Table 1 presents the final test sets.[7] The resulting evaluation suite is a unique collection in terms of the number of dialects, genre and coarse to fine-grained dialect categorization. In total, the evaluation set consists of four coarse dialect categorizes; Nile, Levantine (LEV), Gulf, Maghrebi (MGR) and 15 fine-grained subclasses which are further divided into city-level categories. It belongs to several genres; web-blogs, chat, religion, travel and media including movies, music and news programs. The cross-domain presence of a dialect, e.g. Egyptian data available via APT-eg and MADAR, enables testing the robustness of a dialect translation system across several domains of the same dialect. Similarly, the city-level fine-grained subclasses offer interesting exploration on how dialects between cities of a country are related as shown in (Salameh et al., 2018). In this paper, we provide an opportunity to analyze them in relation to machine translation. In addition to the dialect evaluation sets, we provide the MSA data as part of the suite wherever it is available with the original datasets.

3.3 Dialectness Level Analysis

The evaluation suite consists of a variety of dialects and genre. We attempted to compare them based on their complexity and level of dialectness. We define an approximate metric that we call the *Dialectness Level (DL)*. It is defined as the fraction of tokens that do not overlap with MSA tokens. In other words, the number of dialect words unknown to the MSA vocabulary. DL ranges from 0% to 100% where a high value means high-level of dialectness. Note that there are semantic differences between MSA and dialects

[6] https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

[7] see Appendix A.1 for the complete table including statistics of the development sets.

where an identical word has different meanings, however, such cases are rare and looking at the number of non-overlapping words is a good approximation of the level of dialectness.

We calculate DL for each evaluation set by comparing against the 1.2M MSA words extracted from 40M sentences of the MSA data (see the MSA data in Table 2).[8] First, we compare DL for a single dialect present in two different data sources, e.g. Egyptian from APT-eg and MADAR. Figure 1a shows a few examples of such comparisons. We observed that DL greatly varies between data sources and is not dependent on a specific dialect. For instance, MADAR has a higher DL than Media for Lebanese dialect, but a lower DL than QAraC for Qatari dialect. In some cases like Egyptian, we did not see a major difference between DL of two data sources. The varying amounts of DL shows the variability in the level of difficulty and is a desirable property for an evaluation suite. Next we compare the four coarse dialect categories in terms of their DL. Figure 1b shows that the MGR has the highest DL (6%) while Nile has the lowest DL (2%). This implies that MGR is furthest from MSA and may benefit the least from the MSA data resources. We discuss this further in the context of MT performance in Section 5.

4 Evaluation

In this section, we describe experimental setup and present our results using the evaluation suite.

4.1 Training Data and Evaluation Data

Table 2 summarizes the available Arabic-English training data. The only reasonable sized dialectal training data are of Levantine and Egyptian which consist of 136k and 37k parallel sentences respectively. Rest of the dialect data is very small. We additionally use a large MSA-English corpus to explore the usefulness of MSA in translating dialectal Arabic effectively. The MSA-English corpus consists of OPUS (Lison and Tiedemann, 2016), UN (Ziemski et al., 2016), TED (Cettolo, 2016), NEWS, and QED (Guzmán et al., 2013) corpora. In addition to the dialect testsets mentioned in Section 3, we use five MSA-English testsets – one from News domain, news04, and four from TED talks, test11-14 for some selected experiments.

4.2 Training and Model Settings

Training Settings We build models using various training settings. First, we train systems using the available dialect training data and the MSA data listed in Table 2. Second, we apply the fine-tuning strategy which has shown to be effective in domain adaptation (Sajjad et al., 2017b). In a typical fine-tuning scenario, a large base system is first trained on a heterogeneous data with the understanding that the large amount of data helps to learn the language. In the next step, the parameters of the base system are fine-tuned towards in-domain using the in-domain training data. Here, we loosely consider dialects as a different domain of MSA to maximize the benefit of large available MSA-English parallel data and the small amount of available dialectal training data. Lastly, we use back-translation (Sennrich et al., 2016a) to increase the size of the dialectal Arabic-English training data. We train an English-MSA MT system, fine-tune it on dialects and translate English monolingual data to dialectal Arabic. Then, we use this noisy dialect-English data as an additional training data to improve dialectal Arabic to English translation system.

Model Settings We used transformer-based seq2seq model implemented in OpenNMT (Klein et al., 2017). We used default training and decoding settings: 6 encoder and 6 decoder layers, layer size 512, attention heads 8, dropout 0.1, Adam β_1 0.9, β_2 0.998 and batch size 4096 subwords. For fine-tuning, we additionally use a warmup step size of 800 and label smoothing 0.1.[9] We train for 20 epochs and select the best model using the provided development sets. For example, in the case of a dialect specific system say, Egyptian, we choose the model that performs the best on Egyptian development sets. For a system targeting multiple dialects, we choose the model with the best average performance across development sets of all dialects.

[8] The DL value with respect to each evaluation set is provided in Appendix A.1.

[9] https://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model

	Train	Sentences	Tokens (ar/en)
APT	eg	36k	342k/484k
	lev	136k	1.1m/1.5m
Bible	mr	7k	115k/174k
	tn	7k	113k/174k
	msa	7k	114k/174k
MDC	sy	8k	76k/106k
UN	MSA	18.5M	397M/448M
OPUS	MSA	22.5M	120M/149M
TED	MSA	230k	3.3M/4.0M
NEWS	MSA	322k	7.4M/9.5M
QED	MSA	153k	1.0M/1.3M
NEWS	en	600k	–/12M

Table 2: Arabic-English Training Data. Last row shows the monolingual data used for back-translation

Train	APT-eg$_{dev}$	APT-eg$_{test}$
APT-eg	6.2	6.8
MSA	13.4	14.0
→ APT-eg	**18.7**	**19.3**

(a) Egyptian

Train	APT-lev$_{dev}$	APT-lev$_{test}$
APT-lev	16.5	16.4
MSA	11.0	10.8
→ APT-lev	**20.4**	**19.9**

(b) Levantine

Table 3: Training a dialect-specific system from scratch vs. fine-tuning (→) on the MSA system

4.3 Preprocessing and Evaluation

Following Sajjad et al. (2017a), we morphologically segment the training data using Farasa (Abdelali et al., 2016) and created BPE-based (Sennrich et al., 2016b) vocabulary of 32k operations on the segmented Arabic. Although Farasa is limited to segmenting MSA words present in the dialectal Arabic, it helps to reduce data sparseness. For every individual system, we created its own subword vocabulary. In the case of fine-tuning, we used the BPE-vocabulary of the base system. We used Sacrebleu (Post, 2018) to calculate BLEU (Papineni et al., 2002) scores.

4.4 Experiments

In this section, we explore several training settings, such as fine-tuning, data augmentation, and analyze their effect on the quality of dialect translation.

Can we achieve reasonable translation quality using the small amount of dialectal Arabic-English parallel data? We consider two largest available dialectal training data (Levantine and Egyptian) and train dialectal Arabic-English MT systems using them. We evaluate them on their respective evaluation sets. The first row in Table 3 presents the results. As expected, using limited amount of in-domain dialect training data does not result in decent translation performance. We see very poor translation quality when using only 36k Egyptian data for training. However, the use of Levantine data (136k sentences) resulted in respectable translation performance, with above 18 BLEU score on both dev and test sets.

Can we use MSA for the benefit of dialect translation? Given the large amount of available MSA-English parallel data (≈ 40M sentences), and the lexical and orthographic similarities between MSA and dialects, we explore the usefulness of MSA data in translating Arabic dialects. We train a state-of-the-art MSA-English system and translate Egyptian and Levantine evaluation sets (see the MSA system in Table 3). Using the MSA system alone improved the translation by more than ≈7 BLEU points on Egyptian when compared with the system trained using Egyptian only (APT-eg). On Levantine sets, the MSA system achieved 11.0 and 10.8 BLEU points on dev and test sets but it did not outperform the translation system trained on Levantine alone. These results show that while 36k in-domain sentences are too few to achieve good translation quality alone, increasing the amount of in-domain up to 130k resulted in significantly better quality.

To further utilize the benefits of both the MSA data and the available dialect training data, we used the MSA system as a base system and fine-tune it on the available dialect data. The resulting translation systems (see the last row of Table 3) showed substantial improvement in translation quality of both dialects. The case of Egyptian is interesting where MSA→APT-eg exhibits that with using only 36k dialectal training data and a large MSA system, one can improve the translation by more than 12 BLEU points compared to the Egyptian-only system (APT-eg). On Levantine, we see gains of up to 3.9 BLEU points reaching as high as 20.4 on the dev set. The results conclude that **the use of MSA-base system is beneficial in achieving better translation quality**. The large amount of MSA-English parallel data helps the system to effectively learn the Arabic and English language. The fine-tuning step on a dialectal training data optimizes the weights of the network towards the specific dialect. In rest of the experiments in this paper, we report results using the MSA-English MT system as a base system.

Are dialects helpful in translating each other? The provided dialect evaluation sets consist of a diverse set of dialects, several of which have no dialect-English parallel data available for training and fine-tuning. For example, there is no training data available for Omani dialect and Palestinian dialect. Here, we explore the effectiveness of using all the available dialect training data for zero-shot and a few-shot dialect translation. We concatenate the available dialect training data (see Table 2; APT, Bible and MDC) and fine-tune the MSA system using it. Table 4 (→ALLD) presents the results on the provided test sets.[10] We group the results based on the regions. The use of ALLD substantially improve the performance over MSA in most of the cases. The two clear exceptions are a few MADAR test sets and the MSA test sets. The results of MSA testsets are understandable since →ALLD is optimized for dialects. For MADAR, we observed that these test sets vary in their level of dialectness. A system only optimized using a small amount of dialectal data may not be optimal in all cases (see Section 5 for discussion on this).

The results of APT-eg (row 34) and APT-lev (row 15) are directly comparable with the test results in Table 3 where we use only the data of the specific dialect for training and fine-tuning. Compared to →APT-eg/APT-lev, using all the dialect data (→ALLD) improved the translation performance by 3.3 and 1.6 points on Egyptian and Levantine tests respectively. In case of dialects with no training data such as Palestinian and Qatari, using data of other dialects improve the performance by more than 5 BLEU points in comparison to MSA. These results demonstrate that the data from different dialects is helpful in all scenarios; zero-shot, a few-shot and when as much as 136k parallel sentences are available for training.

Combining dialectal and MSA training data By fine-tuning an MSA system on a limited amount of dialectal training data, the resulting system is optimized to translate dialects present during training and it may be sub-optimal for other dialects. In addition, the limited dialectal training data may result in overfitting. To tackle this, we build upon our observation that in addition to dialectal training data, MSA data is also helpful in translating dialects. Therefore, we fine-tune the MSA system using the concatenation of ALLD and a subset of MSA data (TED + News). We did not choose the full MSA data to limit its influence over the scarce dialectal data. The choice of TED and News is based on covering a general domain of MSA text i.e. spoken language and formal news domain. Note that the optimal choice of genre and the amount of MSA data needed in this experiment requires further empirical investigation.

[10] The results on dev can be found in Appendix A.2.

No.	Corpus	Dia.	SC	City	MSA	→ALLD	→ALLD+MSA$_s$	→BT→ALLD+MSA$_s$
1	Media	MGR	ma	-	5.6	9.4	**9.6**	9.1
2	MADAR	MGR	tn	Sfax	10.8	11.0	13.3	**13.8**
3	MDC	MGR	tn	-	8.2	13.7	13.7	**13.9**
4	MADAR	MGR	tn	Tunis	11.4	13.4	15.3	**16.0**
5	MADAR	MGR	dz	Algiers	17.3	17.2	**21.3**	21.2
6	MADAR	MGR	ma	Rabat	14.7	19.2	22.7	**23.1**
7	MADAR	MGR	ly	Tripoli	22.8	22.0	25.6	**25.9**
8	MADAR	MGR	ma	Fes	20.9	24.8	29.0	**29.9**
9	MADAR	MGR	ly	Benghazi	28.4	27.0	31.6	**32.0**
10	Average				15.6	17.5	20.2	**20.5**
11	MDC	LEV	ps	-	8.7	**15.3**	15.2	15
12	Media	LEV	lb	-	13.0	16.2	**16.8**	16.4
13	MDC	LEV	jo	-	10.7	**17.7**	17.1	16.9
14	MDC	LEV	sy	-	11.1	19.6	18.8	**19.9**
15	APT	LEV	lv	-	10.8	21.5	21.5	**21.9**
16	MADAR	LEV	lb	Beirut	17.0	21.0	23.5	**23.7**
17	MADAR	LEV	sy	Damascus	25.9	27.6	32.5	**33.1**
18	MADAR	LEV	ps	Jerusalem	27.0	27.4	**33.6**	33.5
19	MADAR	LEV	sy	Aleppo	26.4	28.7	**33.7**	34.3
20	MADAR	LEV	jo	Salt$	29.6	28.7	34.4	**34.9**
21	MADAR	LEV	jo	Amman	30.0	29.2	34.4	**35.1**
22	Average				30.0	29.2	34.4	**35.1**
23	QAraC	Gulf	qa	-	11.9	15.8	15.6	**16.0**
24	Media	Gulf	om	-	19.5	18.0	19.2	**19.6**
25	MADAR	Gulf	qa	Doha	27.6	23.9	28.7	**29.3**
26	MADAR	Gulf	iq	Basra	27.7	23.0	28.0	**29**
27	MADAR	Gulf	iq	Baghdad	28.3	23.8	28.6	**29.1**
28	MADAR	Gulf	sa	Jeddah	27.4	24.5	**29.4**	29.1
29	MADAR	Gulf	iq	Mosul	30	26.2	30.9	**31.3**
30	MADAR	Gulf	ye	Sana'a	29.9	26.2	31.0	**31.4**
31	MADAR	Gulf	om	Muscat	**39.5**	32.1	38.4	38.9
32	MADAR	Gulf	sa	Riyadh	**40.7**	33.4	39.7	40.2
33	Average				28.3	24.7	29.0	**29.4**
34	APT	Nile	eg	-	14.0	**22.6**	21.8	22.2
35	MADAR	Nile	eg	Aswan	26.3	25.0	29.9	**30.4**
36	MADAR	Nile	eg	Cairo	28.9	27.0	32.7	**32.9**
37	MADAR	Nile	eg	Alexandria	34.5	31.5	38.2	**38.3**
38	MADAR	Nile	sd	Khartoum	36.7	32.9	38.5	**39**
39	Average				28.1	27.8	32.2	**32.6**
40	MDC	MSA	ms	-	16.9	**20.4**	18.6	19.0
41	Media	MSA	ms	-	29.5	25.5	**29.7**	29.2
42	Media	MSA	ms	-	35.4	29.2	35.0	**35.6**
43	MADAR	MSA	ms	-	**43.4**	33.9	41.2	41.1
44	Average				**31.3**	27.3	31.1	31.2

Table 4: Test results using MSA-English system alone, fine-tuned on the concatenation of dialects (ALLD), plus on a subset of MSA data (ALLD+MSA$_s$), and using back-translated data (BT). → represents the fine-tuning step. Dia. refers to dialect region and SC refers to dialect subclasses. The density of the color presents low to high BLEU scores.

	MT Systems			
Testsets	MSA	→ALLD	→ALLD+MSA$_s$	→BT→ALLD+MSA$_s$
D-Dev	19.02	21.62	24.14	24.39
D-Test	23.04	23.22	26.63	26.95
TED-News	39.38	27.08	38.20	38.64
Avg. of Avg.	27.15	23.97	29.66	**29.99**

Table 5: Average results showing the generalization capability of each system across dialects and MSA

Here, we only intend to show the benefit of mixing MSA data with the dialect data in the fine-tuning step.

The column →ALLD+MSA$_s$ in Table 4 presents the results. On average the translation quality improved by 2.7 in MGR, 5.2 in LEV, 4.3 in Gulf, 4.4 in Nile and 3.8 in MSA when compared with the →ALLD system. We observed a major improvement in the MADAR test sets, where it outperformed the previous best runs. Comparing individual runs, we observed a drop in the performance of MDC (row 13, 14) and in one of the MSA test sets (row 40). However, the overall results showed that it is beneficial to use the MSA data as part of fine-tuning.

Effect of Back-translation Back-translation (Sennrich et al., 2016a) has been an effective method to utilize monolingual target language data to improve the performance of MT systems. Here, we exploit it to increase the size of the parallel dialectal Arabic-English training data. We train a state-of-the-art English to Arabic system using the MSA parallel data (see Table 2) and fine-tuned it on the concatenation of parallel dialectal Arabic-English data. We select the News commentary English monolingual data (600k sentences) available via WMT.[11] We translate it into dialectal Arabic and created a dialectal Arabic-English parallel data. To effectively use the data, we fine-tune the MSA Arabic-English system using back-translated data and then fine-tune the resulting system on the ALLD+MSA data. The last column of Table 4 presents the results. On average, the back-translation improves the translation quality for all dialects and MSA. An increase in the size of the back-translated data would further benefit the dialect translation system. Here, we only show the efficacy of back-translation in this context and leave the exploration of data size for future work.

One-hat-fits-all Arabic translation system Since there exist a large number of dialects, it is not practical to build a separate translation system for every dialect. This would also add an additional layer of dialect identification before sending the input text to the right dialect translation system. Additionally, the dialect boundary is sometimes fuzzy and it may not be possible to confidently assign one dialect to a sentence. From our results in Table 4, we observed that the systems fine-tuned on ALLD+MSA$_s$ perform on average better than the ALLD system and they also perform better on the provided MSA test sets. We further evaluate these systems on the standard MSA evaluation sets of TED talks and News, in order to engage the generalization capability of our system across MSA and a variety of dialects. Table 5 shows the results. The performance of the MSA system on the MSA testsets (TED-News) serves as an upper bound. The →BT→ALLD+MSA$_s$ system is only lower by 0.74 points from the MSA system in translating the MSA testsets while outperforming on dialect evaluation sets by a large margin. It is remarkable that a single system performs well across a variety of Arabic dialects and MSA without any explicit information about the dialect of the input text.

5 Discussion and Analysis

Translation quality vs. Dialectness Level We studied the relationship between translation quality and `DL` of a test set. We hypothesized that a high `DL` may result in low translation quality when using an MSA system (Table 4 – MSA Column). The MGR dialect, particularly Tunisian (row 2, 3, 4) and Moroccan (row 1) showed the lowest BLEU scores, while also having a high `DL` score of 7% and 10% respectively. This implies that a high `DL` contributes towards poor translation quality. However, we found a few exceptions to above observation. For example, while the Iraqi testset has `DL`=7%, the MSA system

[11] We used news-commentary-v15.en data from `http://data.statmt.org/news-crawl/en/`.

still achieved a BLEU score of 29.9 (row 30, Table 4). To further probe this, we perform a qualitative analysis of out-of-vocabulary words (OOVs) from both Iraqi and Tunisian sets, since `DL` reflects the OOVs in a set. We found that OOVs in the Tunisian set are mostly foreign words and highly dialectal lexical choices while several of the OOVs in the Iraqi set are phonological and morphological variations of MSA words. The latter cases are split into known MSA subword units and translated correctly by the system. For example, the Iraqi words تغيد "tgyd",[12] غاح "gAH" , اغيد "Agyd" and غحتو "gHtw" are merely phonologically altered MSA words that are pronounced with غ instead of ر. In contrast, Tunisian used سبيطار "sbyTAr" (Turkish Origin), فريت "fryt" (French), باقاج "bAqAj" (French) for ("hospital","fries" and "luggage" respectively. These lexical variations directly impact the translation quality, adversely in neural MT as they are often split into smaller known units that have nothing to do with the original word.

Lexical Choice Errors Although the inclusion of MSA during fine-tuning significantly improved the translation quality, it also caused lexical ambiguity in some cases. We found that the system erred when translating a few words that have differing semantic meaning between a dialect and MSA. For example, the Tunisian word هزني "hzny" ("carry me") and Moroccan word ديني "dyny" ("take me") got translated into their MSA variants, "shake me" and "my debt" respectively. We conjecture that this happened~~s~~ because of two reasons. First, MSA has a significant influence during training because of its relatively larger size. Second, inadequate context in the input text increases the ambiguity for the system. We found this phenomenon to be frequent in the MADAR testset where average sentence length is only 7 words.

6 Related Work

Data Resources Numerous efforts have been made to build content for dialectal Arabic. Zbib et al. (2012) released Egyptian- and Levantine-English data gathered from weblogs and online user groups, translated through Amazon Mechanical Turk. Bouamor et al. (2014) and Bouamor et al. (2018) created multi-parallel data resources covering various fine-grained categories of dialects. Other efforts outside of MT arena to build resources on Arabic dialect include, but are not limited to (Diab et al., 2014; Khalifa et al., 2016; Jarrar et al., 2017; Suwon et al., 2020; Mubarak et al., 2020).

Machine Translation Machine Translation of Arabic dialects got attention for a short while due to BOLT project. Subsequent efforts were carried to improve MSA-to-English systems by appending Dialect-to-MSA module as pre-processing step (Salloum and Habash, 2011; Salloum and Habash, 2013; Zbib et al., 2012; Sajjad et al., 2013; Durrani et al., 2014; Jeblee et al., 2014) or adapting the MSA-to-English systems towards in-domain dialectal data (Sajjad et al., 2016). Salloum et al. (2014) studied the use of sentence level dialect identification in optimizing MT system selection in mixed dialectal scenario. More recently Baniata et al. (2018) used multi-task learning in neural MT with individual encoders for MSA and dialects and a shared decoder. Despite the number of efforts in translating Arabic dialects to MSA, they are limited to a few dialects and the results among various studies are not comparable due to the difference of evaluation sets. In this paper, we provide the first dialectal Arabic-English evaluation suite based on a large number of dialects, covering various genre and varying amount of dialectness level.

7 Conclusion and Future Directions

We presented **AraBench**, an evaluation suite covering 4 coarse and 15 fine-grained, and 25 city-level dialect categories. The evaluation suite is first of its kind that put together a large number of dialects, covering several domains, and with varying level of dialectness. We adapted an industrial scale MSA-English system to train very strong baselines based on fine-tuning, back-translation and data augmentation and did a large scale evaluation on AraBench. We showed that a single general system can be effectively trained to translate both MSA and a large variety of dialects. The evaluation suite enables numerous future research directions in low-resource MT and dialect translation. For example, adapting unsupervised MT methods towards learning to translate related languages and testing the generalization capabilities of existing methods in MT against varying genre and level of dialectness are a few interesting directions to explore.

[12] Buckwalter transliteration and translation are provided for Arabic words.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16, San Diego, California, June. Association for Computational Linguistics.

Kamla Al-Mannai, Hassan Sajjad, Alaa Khader, Fahad Al Obaidli, Preslav Nakov, and Stephan Vogel. 2014. Unsupervised word segmentation improves dialectal Arabic to English machine translation. In *Proceedings of the Workshop of Arabic Natural Language Processing (ANLP)*, October.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *ACL*.

Laith Baniata, Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). In *Computational Intelligence and Neuroscience Computational Intelligence and Neuroscience*, October.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 1240–1245. European Language Resources Association (ELRA), 1.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The madar arabic dialect corpus and lexicon. In *The International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Mauro Cettolo. 2016. An Arabic-Hebrew parallel corpus of TED talks. In *Proceedings of the AMTA Workshop on Semitic Machine Translation (SeMaT)*, Austin, US-TX, November.

Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Heba Elfardy, Nizar Habash, Abdelati Hawwari, Wael Salloum, Pradeep Dasigi, and Ramy Eskander. 2014. Tharwa: A large scale dialectal Arabic - standard Arabic - English lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3782–3789, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Nadir Durrani, Yaser Al-Onaizan, and Abraham Ittycheriah. 2014. Improving egyptian-to-english SMT by mapping egyptian into MSA. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing - 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II*, volume 8404 of *Lecture Notes in Computer Science*, pages 271–282. Springer.

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a tv broadcasts speech recognition system for qatari arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 3057–3061. European Language Resources Association (ELRA).

Francisco Guzmán, Hassan Sajjad, Stephan Vogel, and Ahmed Abdelali. 2013. The AMARA corpus: Building resources for translating the web's educational content. In *Proceedings of the 10th International Workshop on Spoken Language Technology (IWSLT-13)*, December.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, November. Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the palestinian arabic dialect. *Lang. Resour. Evaluation*, 51(3):745–775.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into Egyptian Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 196–206, Doha, Qatar, October. Association for Computational Linguistics.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. *ArXiv*, abs/1711.00043.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), may.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2020. Constructing a bilingual corpus of parallel tweets. In *Proceedings of 13th Workshop on Building and Using Comparable Corpora (BUCC)*, Marseille, France.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Morristown, NJ, USA.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6.

Hassan Sajjad, Nadir Durrani, Francisco Guzmán, Preslav Nakov, Ahmed Abdelali, Stephan Vogel, Wael Salloum, Ahmed El Kholy, and Nizar Habash. 2016. Egyptian arabic to english statistical machine translation system for NIST openmt'2015. *CoRR*, abs/1606.05759.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, Ahmed Abdelali, Yonatan Belinkov, and Stephan Vogel. 2017a. Challenging language-dependent segmentation for Arabic: An application to machine translation and part-of-speech tagging. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, July.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017b. Neural machine translation training in a multi-domain scenario. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, December.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 767–777.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland, July. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English machine translation: Pivoting through modern standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358, Atlanta, Georgia, June. Association for Computational Linguistics.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence level dialect identification for machine translation system selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778, Baltimore, Maryland, June. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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, August. Association for Computational Linguistics.

Shon Suwon, Ali Ahmed, Samih Younes, Mubarak Hamdy, and Glass James. 2020. ADI17: A Fine-Grained Arabic Dialect Identification Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F 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. Association for Computational Linguistics.

Michal 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 2016, Portorož, Slovenia, May 23-28, 2016.*

A Appendix

A.1 Datasets Statistics

In Table 6, we provide details for both the development and test sets in terms of tokens and types per each dialects to city level.

Corpus	Genre	Dialect	SC	City	dev					test					
					lines	ar		en		lines	ar		en		DL
						tokens	types	tokens	types		tokens	types	tokens	types	
Bible	Religion	MGR	ma	ma	600	9553	3729	14037	2683	600	9427	3559	13808	2640	9.988
		MGR	tn	tn	600	9315	3784	14037	2683	600	9312	3747	13808	2640	6.217
		MSA	msa	msa	600	9319	4300	14037	2683	600	9061	4318	13808	2640	0.752
		MSA	msa	mss	600	7844	3633	14037	2683	600	7545	3535	13808	2640	0.850
MDC	Web blogs	Nile	eg	eg	1000	9557	4797	13470	3105	1000	9220	4581	13052	2964	2.879
		LEV	sy	sy	1000	10153	4396	13470	3105	1000	9611	4167	13052	2964	2.282
		LEV	jo	jo						1000	8167	3675	13052	2964	2.928
		LEV	ps	ps						1000	8506	4054	13052	2964	2.270
		MGR	tn	tn						1000	8746	3844	13052	2964	5.992
		MSA	msa	msa						1000	9983	4357	13052	2964	0.461
APT	Web blogs	LEV	lv	lv	1000	8214	4277	11258	2732	1000	8333	4405	11645	2783	4.253
MADAR	Travel	Gulf	qa	qa	5500	29200	8178	42413	6259	6500	34422	9426	49790	7201	1.762
		LEV	lb	lb	5500	31963	9127	42413	6259	6500	35137	11098	49790	7201	6.568
		MGR	ma	ma	5500	33334	9920	42413	6259	6500	38599	11462	49790	7201	7.889
		MGR	tn	tn	5500	30016	9298	42413	6259	6500	35600	11263	49790	7201	7.133
		MSA	msa	msa	5500	43962	8075	42413	6259	6500	51695	9249	49790	7201	0.089
		Nile	eg	eg	5500	35753	8741	42413	6259	6500	38304	10942	49790	7201	2.197
		Gulf	iq	iq						2000	10771	4719	15262	3492	2.069
		Gulf	om	om						2000	11652	4857	15262	3492	1.164
		Gulf	sa	sa						2000	11102	4457	15262	3492	0.336
		Gulf	ye	ye						2000	11268	4626	15262	3492	2.964
		Gulf	iq	iq						2000	10307	4600	15262	3492	2.376
		Gulf	sa	sa						2000	10581	4398	15262	3492	1.448
		Gulf	iq	iq						2000	11249	4712	15262	3492	7.380
		LEV	jo	jo						2000	11745	4518	15262	3492	1.376
		LEV	pa	pa						2000	10996	4249	15262	3492	2.068
		LEV	sy	sy						2000	10636	4516	15262	3492	3.269
		LEV	jo	jo						2000	12490	4197	15262	3492	1.400
		LEV	sy	sy						2000	10579	4523	15262	3492	3.778
		MGR	dz	dz						2000	11642	4524	15262	3492	4.231
		MGR	ly	ly						2000	11537	4367	15262	3492	3.355
		MGR	ly	ly						2000	11567	4375	15262	3492	3.564
		MGR	ma	ma						2000	11690	4853	15262	3492	5.365
		MGR	tn	tn						2000	10639	4400	15262	3492	7.280
		Nile	sd	sd						2000	11808	4487	15262	3492	1.007
		Nile	eg	eg						2000	11670	4362	15262	3492	1.570
		Nile	eg	eg						2000	11998	4618	15262	3492	2.294
QAraC	TV Show	Gulf	qa	qa	6713	43262	12001	56894	7716	6713	43842	12093	57833	7776	4.747
Media	Movie	MGR	ma	ma						526	8282	3245	8802	2239	10.031
	News	MSA	msa	msa						637	8932	4062	11817	3222	0.511
	News	MSA	msa	msa						621	9282	4043	11107	3009	0.528
	Art show	LEV	lb	lb						250	3901	1846	4952	1522	3.019
	Cultural Prg.	Gulf	om	om						467	6520	3080	7976	2441	0.726

Table 6: Development and test sets details for both Arabic (ar) and English (en) parallel sides in terms of tokens and types and DL levels per dialect.

A.2 Machine Translation Results

Table 7 presents the results of the provided development sets. In addition, we provided the results of the Bible test sets in Table 8.

Corpus	Dia.	SC	City	MSA	ALLD	→ALLD+MSA$_s$	→BT→ALLD+MSA$_s$
MDC	LEV	sy	-	9.2	17.9	17.8	18.2
APT	Nile	eg	-	13.4	21.1	21.1	21
APT	LEV	-	-	11.0	22.6	21.9	22.5
MADAR	Gulf	qa	Doha	28.1	24	29.2	29.8
MADAR	LEV	lb	Beirut	21.8	25.4	28.8	29.3
MADAR	MGR	ma	Rabat	10.0	14.3	16.4	16.6
MADAR	MGR	tn	Tunis	12.1	14.1	16.4	16.6
MADAR	MSA	ms	-	45.8	35.4	43.3	42.9
MADAR	Nile	eg	Cairo	27.1	25.7	30.8	31.1
QAraC	Gulf	qa	-	11.7	15.7	15.7	15.9
Bible	MGR	ma	-	4.1	24.5	25.7	27.1
Bible	MGR	tn	-	7.2	25.9	26.5	27.5
Bible	MSA	ms	-	16.6	31.3	31.8	33.5
Bible	MSA	ms	-	12.9	26.1	27.0	29.0

Table 7: Machine Translation Results on the development sets

Corpus	Dia.	SC	City	MSA	ALLD	→ALLD+MSA$_s$	→BT→ALLD+MSA$_s$
Bible	MGR	ma	-	4.2	26.6	27.8	28.8
Bible	MGR	tn	-	7.0	26.4	27.7	29.2
Bible	MSA	ms	-	17.0	30.8	31.8	33.2
Bible	MSA	ms	-	12.8	27.3	28.4	29.2

Table 8: Machine Translation Results on the Bible test sets