The SADID Evaluation Datasets for Low-Resource Spoken Language Machine Translation of Arabic Dialects

Wael Abid
Department of Computer Science
Stanford University
wael@cs.stanford.edu

Abstract

Low-resource Machine Translation recently gained a lot of popularity, and for certain languages, it has made great strides. However, it is still difficult to track progress in other languages for which there is no publicly available evaluation data. In this paper, we introduce benchmark datasets for Arabic and its dialects. We describe our design process and motivations and analyze the datasets to understand their resulting properties. Numerous successful attempts use large monolingual corpora to augment low-resource pairs. We try to approach augmentation differently and investigate whether it is possible to improve MT models without any external sources of data. We accomplish this by bootstrapping existing parallel sentences and complement this with multilingual training to achieve strong baselines.

1 Introduction

Machine Translation (MT) models have achieved state-of-the-art results on several high-resource language pairs. This success is mostly due to advances in modeling and the availability of clean parallel corpora. Yet, advances in MT on low-resource language pairs are still lagging behind mainly due to the scarcity of parallel training data. Augmenting with back-translated large monolingual corpora (Sennrich et al., 2016a) can partly offset the effects of small amounts of training data, but it is not applicable for pairs where the target language doesn't have large monolingual corpora. On the other hand, freely available benchmark datasets across different tasks have historically provided reference points for researchers to drive their fields forward. Notable examples include The General Language Understanding Evaluation (GLUE) (Wang et al., 2018) for NLU and the Conference on Machine Translation (WMT) datasets for MT. Still, benchmark datasets for most low-resource language pairs remain a much-needed resource.

Arabic dialects, like most low-resource languages, lack freely available benchmark datasets that can be used to evaluate models. This makes previous research results difficult to track and reproduce.

In this paper, we introduce benchmark datasets for evaluation on MT tasks between Arabic dialects, Modern Standard Arabic (MSA) and English. We describe the design considerations and data collection guidelines we adopt. We provide an analysis and an empirical evaluation of state-of-the-art MT models on our benchmark datasets. We explore optimal unsupervised segmentation parameters and introduce a novel data augmentation method which we call bootstrapping.

The goal of the project is to provide datasets that are reliable public evaluation benchmarks to track progress in the translation quality across different dialects. We put a specific emphasis on creating datasets that are not only able to assess the performance of MT systems but also test their robustness on two levels: the domain, and the dialectal diversity.

2 Related Work

Machine Translation resources for Arabic are mostly focused on MSA. Nonetheless, there are a number of efforts dedicated to dialects. The BOLT (Broad Operational Language Translation)[1] program

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

[1] https://www.ldc.upenn.edu/collaborations/current-projects/bolt

Proceedings of the 28th International Conference on Computational Linguistics, pages 6030–6043
Barcelona, Spain (Online), December 8-13, 2020

includes parallel training resources for Egyptian and Levantine dialects. These resources, however, are only available through a subscription from the Linguistic Data Consortium. Bouamor et al. (2014) presented a small scale parallel 7-way corpus composed of five dialects, MSA and English. Translations were produced from Egyptian sentences, which resulted in translations being biased by some Egyptian expressions. PADIC (Meftouh et al., 2015) is a 6-way parallel corpus that includes five dialects and MSA, but not English. It has the same issue as Bouamor et al. (2014) since translations were made from MSA and dialect source sentences. Bouamor et al. (2019) introduce datasets for city-level Arabic dialects for the dialect identification task. The dataset is translated from the English Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007), a corpus composed of short sentences in the travel domain. Unfortunately, while the Arabic translations are available, the English sentences are not.

On the modeling side, several efforts to improve MT systems have been proved to work well for the low-resource task. Sennrich et al. (2016a) made use of monolingual data by augmenting the parallel text with back-translated sentences. Lample et al. (2018) leveraged back-translation and the denoising effect of a language model in order to generate synthetic parallel data while having access to monolingual data only. Additionally, Zoph et al. (2016) made use of transfer learning by training on a related high-resource pair and fine-tuning on the low-resource pair. In a similar approach, Liu et al. (2020) demonstrated that multilingual denoising pretraining produces significant performance gains in low-resource tasks. Fadaee et al. (2017) proposed an approach that augments the training data by generating new sentence pairs with synthetically created contexts targeting low-frequency words.

3 Arabic and its Dialects

Arabic is an example of the linguistic phenomenon diglossia (Ferguson, 1959), where its written format, Modern Standard Arabic (MSA), differs greatly from the regional spoken dialects. Mutual intelligibility between dialects is limited. For example, most Arabic speakers understand Egyptian due to the popularity of Egyptian movies and television shows across the Arab world. Though, the Tunisian dialect is quite difficult to understand for an Egyptian speaker. At a high level, Habash (2010) divides Arabic dialects into 5 regional clusters: Maghrebi Arabic (Mauritania, Morocco, Algeria, Tunisia, and Libya), Egyptian Arabic (Egypt and Sudan), Levantine Arabic (Lebanon, Syria, Jordan, and Palestine), Gulf Arabic (Qatar, Kuwait, Saudi Arabia, United Arab Emirates, Iraq, and Bahrain) and Yemeni Arabic. However, dialects differ from country to country and even from city to city.

MSA refers to the formal Arabic that developed in the 19th century from Classical Arabic, the language of the Quran and the literature starting from the 7th century. Differently, each Arabic dialect is the product of the pre-Arabization language of the corresponding geographical location, and other historical factors like colonization. For example, Arabic dialects in North Africa use a number of French loanwords and still hold some words from the native languages of the Amazigh aboriginals.

MSA is the only variety that is standardized, taught in schools, and used in the media and in official documents. The dialects are not taught in schools, and they are mainly used for day-to-day conversations. Moreover, until the advent of social media, they have remained rather absent from written formats. This, in addition to the MSA's prevalence in written form, explains why almost all Arabic datasets have predominantly MSA content. Unlike MSA, dialects don't have an established orthography. However, it is possible to write Arabic dialect text by using the spelling rules of MSA, which are mostly phonetic.

In essence, Arabic dialects not only introduce a low-resource challenge, but they also present the challenge of a spoken language that has phonological, morphological, syntactic, lexical, and orthographical variability with every speaker even within the same geographical location. These factors need to be accounted for in order to create a reliable evaluation benchmark for Arabic dialects.

4 Dataset Creation

Arabic dialects are spoken languages and they are mainly used for day-to-day communication. When it comes to less ordinary topics, people tend to use loanwords from another language (e.g. MSA, English, etc.) or switch to another language. We wanted the domain of our datasets to be an accurate representation of the topics people discuss when speaking their dialects. Therefore, we conducted a survey where we asked native speakers the following questions (a) In general terms, what are some topics of your conversations when you're speaking your dialect? and (b) In general terms, what are some topics where it's more convenient to switch to another language or borrow terms from another language?

We obtained responses from 65 native speakers from 6 Arabic speaking countries (Morocco, Algeria, Tunisia, Egypt, Palestine, Syria), which we summarize in Figure 1. Results of the survey indicate that people use their dialects to discuss diverse topics touching upon daily life, politics to culture. Topics, where dialects are typically not used, mainly relate to highly technical concepts. We summarize our design motivations in (i) accurately reflecting the topics that people discuss when speaking their dialects, and (ii) capturing the dialectal diversity within the same dialect. We create our dataset in three stages: selecting and curating source passages, translating the resulting passages, and performing quality checks.

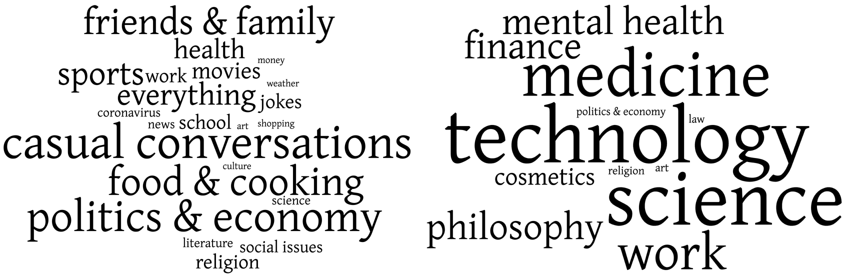

Figure 1: Word clouds summarizing the results of questions (a), left and (b), right.

4.1 Selecting and curating source passages

In order to mimic such topics from the survey, we picked our documents from three sources: Simple Wikipedia for its domain diversity and straightforward language, Aesop Fables for its anecdotal style, and from select conversations from movie subtitles. We chose English as the language of our source sentences instead of MSA as not to bias our translations (Bouamor et al., 2014). We start with a Simple Wikipedia snapshot from April 2020 containing 160,935 articles and 664,603 sentences. In order to exclude the undesired topics, we tried filtering out sentences that have technical terms. Our approach was to automatically filter out sentences that have words with a high tf-idf score. This indicates that these words are more relevant to those sentences than others and are likely technical terms. We also filtered out sentences that have less than 8 words, and that have characters other than the alphabet, punctuation, numbers, and spaces. We were left with a small 2,365-article, 5,263-sentence subsample of the dump which we went through manually to take out more sentences with technical terms. For the Movie Subtitles, we started with the Cornell Movie Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011) which is composed of a set of 83,097 dialogs from a number of movie subtitles. We filtered out dialogs that have less than 6 sentences and those that contain profanity. Our final step was to review the Aesop Fables to slightly simplify the sentence structure to make it more converstational in style. Examples of the transformations we made are in Appendix A.

For each of the three sources, we chose to pick sentences from as many documents as we can to diversify the topics as much as possible. This diversity of topics increases the difficulty of the datasets. This makes them not only useful for testing the quality of translations but also for testing the robustness of the MT system and the effects of domain drift, making it a solid testbed for multiple translation experiments. Statistics about the resulting English source sentences are included in Table 4 of Appendix B.

4.2 Translation

Different translators provide translations with different levels of dialectness (Habash et al., 2008). Additionally, a dialect, as well as its spelling, could vary from a speaker to another. We think that reflecting this diversity in the datasets would ensure that the datasets are able to evaluate another level of robustness with respect to orthography and vocabulary. In order to create benchmark datasets for Egyptian, Levantine, and MSA, we hired 3 professional translators per language. The translators, who usually translate between English and MSA, are native speakers of their dialects and they originate from different cities. The Egyptian translators are from Cairo, Alexandria, and Qena. The Levantine translators are from Nablus, Gaza, and Khan Yunis. The lack of standardized orthography rules for Arabic dialects can cause one word to be spelled in different ways by each native speaker. This would result in a very sparse vocabulary. In order to reduce this effect and ensure uniform translation quality across the samples, the translators were presented with guidelines similar to those in (Bouamor et al., 2019). Each translator was given an equal amount of sentences from the total source dataset with the goal of creating three splits similarly to (Guzmán et al., 2019). The resulting translations are a *dev* set for hyperparameter tuning and model selection, a *devtest* set for measuring generalization during development, and a blind *test* set for the final evaluation.

4.3 Quality evaluation

While we made sure to prepare translatable English source sentences and provided the translators with clear guidelines, translating into an Arabic dialect is neither a common nor a simple task. To ensure the quality of translations, we added one last quality check step to our data creation process.

For each dialect, we asked 2 independent native speakers to rate the translations on two criteria each on a scale of 1 to 20: the preservation of meaning between the source and the translation, and the naturalness of the translation. Sentences getting under 15 in either of the criteria from either of the raters are sent for rework. Once re-translated, sentences are evaluated again and only those that get 15 or more in both criteria from both raters are kept.

5 Resulting dataset

We obtain a 4-way dataset between English, MSA, Egyptian, and Levantine that can be used in several different MT and Dialect Identification experiments. For each language, the dataset sizes for the *dev*, *devtest*, and *test* sets are 2,997, 2,997, and 2,994 sentences respectively.

To understand the properties of our datasets and quantify the similarities between the languages, we compute the Jaccard Similarity Coefficient between pairs of the three Arabic languages. We see that there's a considerable 0.33 similarity between the two dialects. We find that MSA is slightly closer to Levantine (0.30) than it is to Egyptian (0.29), and that even an intersection over the three languages still shows high lexical similarity (0.18). This hints that a multilingual setting could be helpful. Further statistics about the different splits are included in tables 5 and 6 in Appendix C.

6 Experiments

6.1 Data

To the best of our knowledge, the only data publicly available for Egyptian (EGY) and Levantine (LEV) is distributed through the Linguistic Data Consortium[2]. For Egyptian, we use data from catalogs LDC2012T09 (Zbib et al., 2012), LDC2019T01, LDC2019T18, and LDC2020T05. For Levantine, we use LDC2012T09. These catalogs are part of the BOLT project and they've been translated from dialectal text retrieved from forums, transcribed phone conversations, SMS, and chats. For our MSA and multilingual experiments, we also use a selection of parallel MSA data from news and web blogs translated as part of the GALE (Global Autonomous Language Exploitation) program (Cohen, 2007). We learn a joint source and target Byte-Pair-Encoding (Sennrich et al., 2016b) which will be discussed more in detail in section 6.3. Our total data amounts to 402k sentences and 3.80M tokens for Egyptian, 138k sentences and 1.27M tokens for Levantine, and 1.49M sentences and 24.45M tokens for MSA.

6.2 Models and Architecture

All of our experiments were conducted with the fairseq toolkit (Ott et al., 2019). We used the Transformer (Vaswani et al., 2017) base architecture with 6 encoder and 6 decoder layers, 8 attention heads, an embedding size of 512, an inner-layer dimension of 2048. We use the Adam optimizer (Kingma and Ba,

[2] ldc.upenn.edu

2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$. We chose a learning rate of $\epsilon = 5\text{e-}4$ with weight decay and 4000 warmup steps. For regularization, we use a label smoothing rate of 0.2 and a dropout rate of 0.4. We report performance with BLEU (Papineni et al., 2002) using `multi-bleu.perl`[3]. Our hyperparameters were tuned using the *dev* set.

6.3 Training setup

Careful BPE selection

Translation is an open-vocabulary problem, and using Byte-Pair Encoding (BPE) (Gage, 1994) helps solve the reliance of NMT systems on a fixed vocabulary. Sennrich et al. (2016b) first used BPE for MT based on the intuition that various word classes are translatable via subwords. This is especially useful for languages with agglutination and rich morphology. Arabic dialects' lack of orthography rules and their morphological richness could make it difficult to learn an accurate mapping between languages. This challenge is showcased through our EGY side and ENG side training data vocabulary sizes of 287k and 116k respectively, and of 137k and 61k for the LEV side and ENG side (note that the LEV-ENG pair data amount is one-third that of the EGY-ENG pair). This, along with the low-resource nature of these languages, calls for carefulness when choosing the BPE vocabulary size. Therefore, we conduct a set of experiments where we search for an optimal BPE vocabulary size for both Egyptian and Levantine dialects, and we examine whether the translation direction has an effect on the quality.

Bootstrapping

Arabic dialects are not official languages of their countries and are mainly only spoken. This, along with the dominance of MSA in written formats, prevents dialects from having large monolingual corpora.
Consider that we want to improve an MT system in the English-Dialect direction. As the Arabic dialect is on the target side, back-translation isn't a possibility since we lack monolingual data. We introduce bootstrapping, illustrated in Figure 2, a simple but efficient idea used to augment the training corpus.

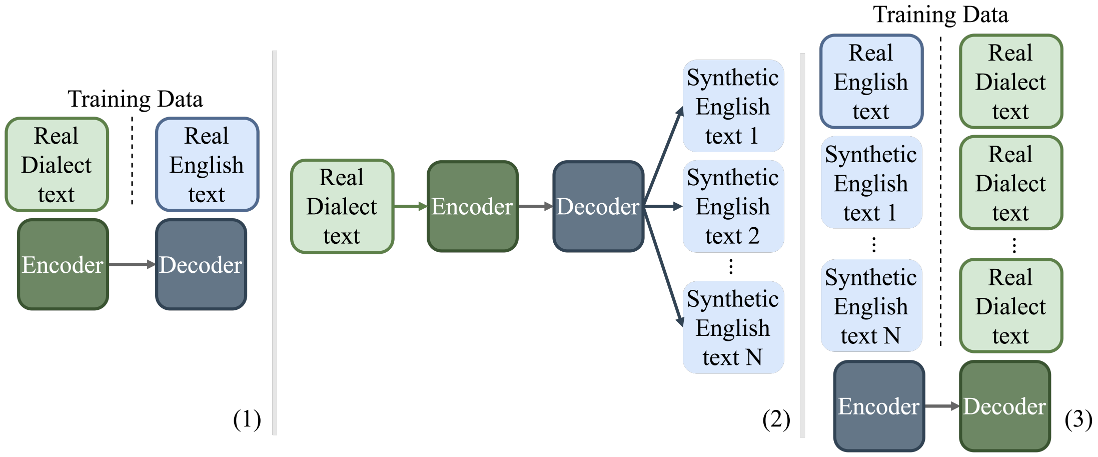

Figure 2: Bootstrapping in three steps: (1) Train the reverse Dialect-English MT system (2) Use this system to back-translate the training corpus several times, while varying decoding parameters each time in order to generate diverse candidates (3) Once we have our bootstrapped sentences, we append them to our original training data and train the English-Dialect MT model.

This way, we are able to generate diverse candidates of the same sentence and duplicate our corpus by sampling from the same training distribution, hence the name bootstrapping. We investigate two methods for generating synthetic diverse source sentences given our backward MT system. The first is beam search with beam size 5 and softmax temperature and length penalty variation at the generation step, and the second is sampling.
Beam search optimizes for the output with the highest probability, therefore leading to synthetic

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

sentences that are rather similar. We try to offset this effect by jointly varying two parameters. Each time we generate synthetic sentences from the training data, we set a different softmax temperature, sometimes using a high temperature in order to get diversified outputs, and sometimes using a low temperature to get more similar outputs for an overall diversity. At the same time, we vary the length penalty favoring long sentences at times, and shorter sentences at other times.

The sampling method, unlike beam search, samples the tokens from the probability distribution instead of finding the argmax. This makes it a better approximation of the data distribution and therefore allows us to get more diverse outputs. We decide to control the level of randomness that sampling offers by restricting it to the top 10 most probable words, so that the generated sentences are not too divergent. Table 7 in Appendix D is an example of the relative richness of outputs of top 10 sampling versus beam search with various temperature and length penalty parameters.

Multilingual Training

Johnson et al. (2017) and Firat et al. (2016) showed that multilingual NMT models trained on large amounts of data outperform bilingual models in most cases, and that multilingual models are particularly useful for low-resource settings. In a multilingual model, some or all parameters are shared by all the language pairs it is trained on. This creates a shared semantic space between languages which benefits the model's ability to generalize well, and allows positive transferability from high-resource to low-resource languages. Furthermore, a multilingual MT model's vocabulary is the union of the vocabularies of all languages, which could include significant intersections if subword units are used and if languages share the same character set. In our case, performing an intersection over the languages' vocabulary set and computing the Jaccard similarity score between Egyptian, Levantine and MSA showed that there's significant overlap between the languages. This overlap is a good base for a multilingual training setting, and the resulting shared semantic representations could be beneficial for the low-resource pairs.

7 Results

In this section, we analyze our results from Table 1. Our analysis first examines different BPE vocabulary sizes in order to recommend an optimal hyperparameter for the language pairs. Next, we compare the two different generation methods used for bootstrapping and contrast it with back-translation. We also analyze the effects of multilingualism and the role of MSA in multilingual training.

	EGY-ENG	ENG-EGY	LEV-ENG	ENG-LEV	MSA-ENG	ENG-MSA
BPE 1,000	18.09	3.74	12.66	2.36	-	-
BPE 5,000	18.17	3.90	12.31	2.36	27.56	15.66
BPE 7,500	18.10	3.76	11.91	2.25	-	-
BPE 10,000	18.10	3.84	11.48	2.20	-	-
BPE 20,000	17.53	3.68	10.49	2.12	-	-
2x Beam Bootstrap	-	4.06	-	2.47	-	-
5x Beam Bootstrap	-	4.12	-	2.51	-	-
10x Beam Bootstrap	-	4.21	-	2.55	-	-
2x Sampling Bootstrap	-	4.08	-	2.54	-	-
5x Sampling Bootstrap	-	4.24	-	2.64	-	-
10x Sampling Bootstrap	18.93	4.36	13.77	2.71	-	-
Back-translation	19.52	-	14.94	-	-	-
Dialect-only Multilingual (no MSA)	17.75	3.22	17.25	3.68	-	-
Multilingual	21.93	4.11	20.89	4.44	30.12	11.23
Multilingual + Sampling Bootstrap	22.13	4.19	21.60	4.66	30.74	11.64

Table 1: Translation performance (BLEU) on the *test* set for all experiments.

Overall scores are rather low, meaning that this is a challenging benchmark for future experiments. Unsurprisingly, translating into the more morphologically rich Arabic dialects produces lower scores. We see that the scores for the MSA-ENG pair are significantly higher than those of the dialects. We attribute this to the fact that MSA is a high-resource language and that it doesn't have the linguistic irregularities that characterize dialects.

7.1 Vocabulary size for Byte-Pair Encoding

We train the architecture described in §6.2 on our parallel data for both EGY-ENG and LEV-ENG pairs in both directions. In each experiment, we vary the BPE vocabulary size from 1,000 to 20,000 and train until convergence.

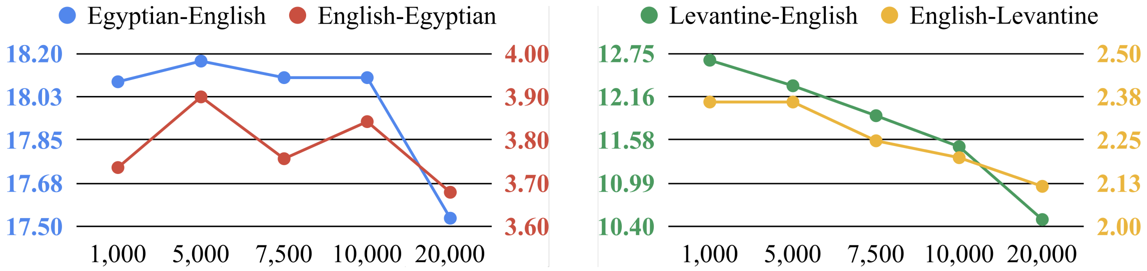

Figure 3: Effect of varying Byte-Pair Encoding vocabulary size on BLEU score.

A BPE vocabulary size in the range of 30k-40k is common in the literature (Ding et al., 2019). Though, Figure 3 shows a trend that indicates that this vocabulary size isn't suitable for either of these pairs. This is potentially the case for many other low-resource pairs, meaning that this exploration step is detrimental to the quality of the MT system. Figure 3 indicates that bigger BPE vocabulary sizes get lower BLEU scores. A potential hypothesis that could explain this is that the combination of a small dataset and a disproportionately large vocabulary causes a lot of sparsity in the vocabulary which hurts the model's performance.

For both directions of the EGY-ENG pair, a vocabulary size of 5,000 is the optimal hyperparameter. However, for both directions of the LEV-ENG pair, a smaller vocabulary size of 1,000 yields the highest BLEU scores. This difference potentially is caused by the smaller size of the LEV-ENG pair training data, amounting to one-third of EGY-ENG data, which could exacerbate the irregularities of the dialect.

Beam search vs. Sampling for Bootstrapping

For each of the ENG-EGY and ENG-LEV directions, we train MT systems on the concatenation of original parallel data and bootstrapped sentences. 10x bootstrap means that we augmented the original data with a synthetic amount equal to 9x the parallel sentences.

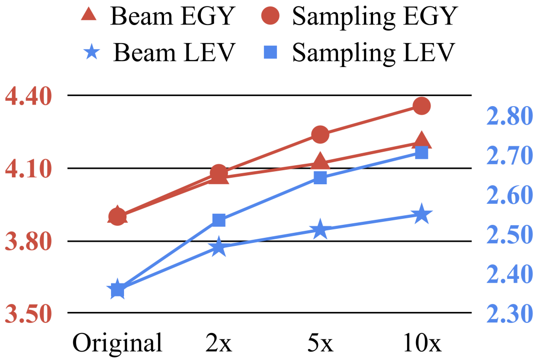

Figure 4: Effects of bootstrapping with Beam search vs. Sampling and the amount of bootstrapped data.

Figure 4 shows immediate improvement over the original data size, even in the smallest data setting.

We observe that bootstrapping using sampling improves both ENG-EGY and ENG-LEV by 0.46 (11%) and 0.35 (14%) BLEU respectively and outperforms bootstrapping with beam search by 0.15 and 0.16 BLEU for the same pairs in the largest data setting. This difference illustrates that lexical diversity from sampling enriches the training and that the bootstrapped synthetic sentences provide useful information that complements our original parallel data.

In order to compare bootstrapping with back-translation on EGY-ENG and LEV-ENG, we train two 10x sampling bootstrap systems and two other systems on data augmented with back-translated sentences, with back-translation performed in the ENG-EGY/LEV direction. Both systems are trained on the same amount of data. Compared with the corresponding baselines, BLEU scores for systems trained with bootstrapping vs. those trained with back-translation are respectively 18.93 (+0.76) and 19.52 (+1.35) BLEU in the EGY-ENG direction and 13.77 (+1.11) and 14.94 (+2.28) BLEU in the LEV-ENG direction. While back-translation offers double the improvement of bootstrapping, it remains a method that requires external resources that are sometimes lacked. Bootstrapping, on the other hand, is a language-agnostic method that doesn't require any external data, yet still provides considerable improvements.

7.2 Multilingual Training

In this section, we analyze our results from experiments on multilingual training and we investigate the role of MSA in this setting. For each of the Dialects-English and English-Dialects directions, we start by training a multilingual model with Egyptian and Levantine dialects only. In a separate experiment, we add MSA to the training. Lastly, we add 10x sampling bootstrapped data to the multilingual + MSA setting for both directions. The architecture is the same as the previous experiments, except for sharing the encoder in the Dialects-English direction, and sharing the decoder in the English-Dialects direction.

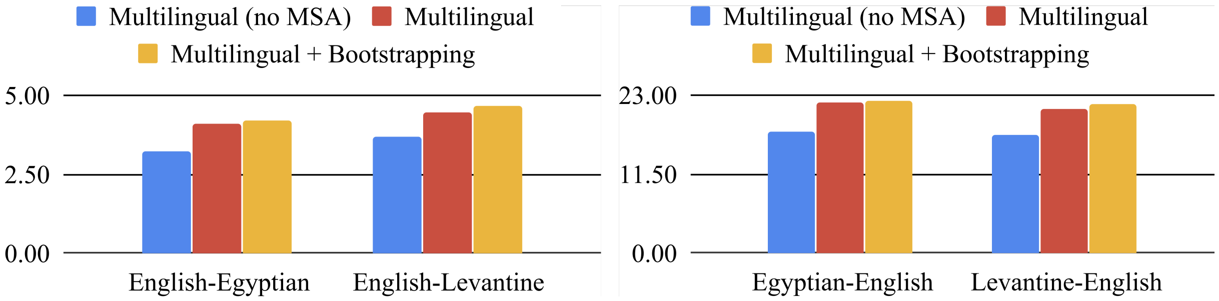

Figure 5: Effects of adding MSA and using bootstrapping in a multilingual setting.

Despite a significant lexical similarity between Egyptian and Levantine, dialect-only multilingual models were only beneficial for Levantine and not for Egyptian. Compared to the baseline, we see an improvement of 4.59 (36%) and 1.36 (55%) BLEU for LEV-ENG and ENG-LEV respectively, but we see decreases of 0.42 (2%) and 0.68 (21%) BLEU for EGY-ENG and ENG-EGY respectively. The effect of MSA in improving the translation quality of the dialects is salient. Compared to the dialect-only multilingual system, adding MSA results in improvements of 3.76-4.18 (21-23%) BLEU in the Dialect-English direction and 0.76-0.89 (20-27%) BLEU in the English-Dialect direction. On the other hand, the effect of dialects on MSA translation quality is also visible, as the dialects cause a decrease of 4.43 (39%) BLEU in the ENG-MSA direction compared to the baseline. Nonetheless, the multilingual model outperforms the baseline in the MSA-ENG direction, improving the score by 2.56 (9%) BLEU.

The results from the last two experiments accurately describe the relationship between MSA and Arabic dialects. Dialects inherit a lot from MSA but not the opposite, which explains why MSA provides a significant performance boost to the dialects but not vice versa.

Even in the high-resource multilingual setting, bootstrapping is still able to improve BLEU score by 0.08-0.22 (1-4%) in the English-Dialect direction. These margins are not as important as those from previous bootstrapping experiments, and this is perhaps due to the fact that other languages in the multilingual setting are able to provide part of the lexical diversity that bootstrapping provides.

8 Discussion

8.1 Source diversity and domain drift

We would like to understand the effect that the diversification of data sources has on evaluation, and how baseline, bootstrap, and back-translation models perform on each of the sources. We therefore separate our *test* set into its three sources, movie subtitles, Aesop Fables, and Wikipedia, respectively amounting to 886, 833, and 1275 sentences and test each subset on the three LEV-ENG systems.

	Movie Subtitles	**Aesop Fables**	**Wikipedia**	**Full *test* set**
BPE 1,000	11.66	11.32	13.99	12.66
10x Sampling Bootstrap	12.46	11.70	15.78	13.77
Back-translation	11.52	11.04	19.40	14.94

Table 2: Translation performance (BLEU) per domain on three Levantine-English systems.

Table 2 shows that bootstrapping outperforms the baseline translation quality over all subsets in nearly proportional increments, whereas the back-translated data (news and weblogs) significantly improves the system's performance on the Wikipedia subset, but falls short when evaluating on movie subtitles and Aesop Fables compared to the baseline and the bootstrap systems. We attribute the decrease in scores to the fact that we augmented the original parallel sentences with 9x its size worth of back-translated data from a different domain, and we think that oversampling the real LEV-ENG data to an equal amount could fix the drop in performance. The difference in BLEU scores over the three subsets showcases the effects of domain drift. Furthermore, it points out to the effect of source diversification while building the benchmark datasets and its role in testing the robustness of MT systems.

8.2 Translator diversity and dialect differences

In order to quantify the mismatch between the dialects in the training data and the dialects in our dataset, we evaluate the baseline ENG-EGY system on the same sentences from our *dev* and *devtest* sets.

	ENG-EGY (*dev*)	**ENG-EGY (*devtest*)**
Translator 1 (Cairo)	4.49	4.10
Translator 2 (Alexandria)	4.50	4.10
Translator 2 (Qena)	4.35	3.97

Table 3: English-Egyptian Translation performance (BLEU) per translator on the *dev* and *devtest* sets.

Table 3 shows that our system performs worse on the Egyptian dialect from Qena than those from Cairo and Alexandria, suggesting that these city dialects are different enough and that the northern Egyptian dialect is more represented in the training set.

9 Conclusions

Despite remarkable previous efforts and advances, low-resource MT, more specifically for spoken languages, remains a challenge mainly due to the lack of training and evaluation data. In this work, we create and make freely available to the community a 4-way benchmark dataset between Egyptian, Levantine, MSA, and English with the aim of providing a reliable resource for the research community to test their systems. We conduct experiments where we show the value behind our design decisions, and we call on the research community to consider such choices, namely the diversity in the source data as well as the translators, when creating resources for low-resource spoken languages in the future.

We evaluate on a number of state-of-the-art baselines and explore optimal training settings for such language pairs. We achieve a significant improvement over the baseline without external data by using a simple augmentation technique, which we call bootstrapping. Our findings suggest that a multilingual model of dialects and MSA, along with bootstrapping, achieves the best results.

References

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy, August. Association for Computational Linguistics.

J. Cohen. 2007. The gale project: A description and an update. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 237–237.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland, August. European Association for Machine Translation.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Charles A Ferguson. 1959. *Diglossia*, volume 15. Taylor & Francis.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

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.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Faraj. 2008. Guidelines for annotation of arabic dialectness. *Proceedings of the LREC Workshop on HLT NLP within the Arabic world*.

N.Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis digital library of engineering and computer science. Morgan & Claypool Publishers.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2001.08210*.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China, October.

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. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. 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.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324, September.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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, November. Association for Computational Linguistics.

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, Montréal, Canada, June. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, November. Association for Computational Linguistics.

Appendix A

Example 1
Original: “Two Goats, frisking gayly on the rocky steeps of a mountain valley, chanced to meet, one on each side of a deep chasm through which poured a mighty mountain torrent.”
Simplified: “Two Goats who were jumping happily on the rocky mountain valley, met by chance, one on each side of a deep canyon through which poured a mighty mountain river.”

Example 2
Original: ”Away with you, vile insect! Said a Lion angrily to a Gnat that was buzzing around his head.”
Simplified: ”Go away, evil insect! Said a Lion angrily to a Fly that was buzzing around his head.”

Appendix B

Source	Number of sentences	Number of words	Avg. number of words per sentence	Number of documents	Percentage from total
Simple Wikipedia	2723	37550	13.79	958	45.05
Aesop Fables	1647	21427	13.01	147	25.70
Movie Subtitles	1757	24387	13.88	208	29.25
Total	6127	83364	13.60	1351	100

Table 4: Statistics on the English source sentences of the benchmark data.

Appendix C

set	Number of sentences	Number of English words	Number of Egyptian words	Number of Levantine words	Number of MSA words
dev	2,997	40,885	37,480	36,362	37,384
devtest	2,997	41,946	37,928	37,928	37,901
test	2,994	40,587	38,672	37,187	38,512

Table 5: Number of sentences and words per language, for the *dev*, *devtest*, and *test* sets.

Vocab set	English	Egyptian	Levantine	MSA	Egyptian ∩ MSA	Levantine ∩ MSA	Egyptian ∩ Levantine	EGY ∩ LEV ∩ MSA
Vocab size	8,449	18,702	20,269	14,903	7,686 (0.29)	8,197 (0.30)	9,859 (0.33)	6,311 (0.18)

Table 6: Vocabulary size per language and lexical similarity between pairs.

Appendix D

Sentence in Levantine: وتعتبر الزهرة الوطنية للملكة الأردنية الهاشمية

Decoding with beam search	**Decoding with sampling**
It is considered the national flower for the Hashimian queen	It is considered the national flower for the Hashimite Kingdom of Jordan
It is considered the national flower for the Hashimi Kingdom of Jordan	And it is considered the national flower for the Hashimi Kingdom of Jordan
It is considered the national flower for the Hashemite Kingdom of Jordan	And the national flower of Jordanian queen is considered an noble flower
It's considered the national flower for the Hashimi Kingdom of Jordan	It is considered a national rose to the Hashemite Kingdom of Jordan
And the national flower is for the Hashimi Kingdom of Jordan	And it's considered the national Jordanian flower of the Hashimi Kingdom

Table 7: Examples of bootstrapped sentences using beam search and sampling.

Appendix E

Language	Sentences
Egyptian	الحب زي ما هوه متغيرش. الناس هيه اللي بتتغير.
Levantine	الحب هو نفسه ما بتغير. الناس هم اللي بتغيروا.
MSA	الحب هو الحب كما كان دائما. الناس هي التي تتغير.
English	Love’s the same as it always was. It’s people who change.
Egyptian	آه طبعا، أكيد، وأنا أراهن إنك اتعشيت استاكوزا امبارح. مع إزازتين شمبانيا.
Levantine	آه، أكيد، وبراهنك إنك ماكل سرطعون مبارح بالليل. مع قنينتين شامبانيا.
MSA	نعم، بالتأكيد، وأنا أراهن أن تناولت سرطان البحر الليلة الماضية. مع زجاجتين من الشمبانيا.
English	Yeah, sure, and I bet you had lobster last night. Along with two bottles of champagne.
Egyptian	في لمح البصر كان القط فوق شجرة، مستخبي بين ورق الشجر.
Levantine	في لحظة البسة كانت على ظهر الشجرة، متخبية بين الأوراق.
MSA	في لمح البصر كان القط فوق شجرة، مختبئا بين الأوراق.
English	In an instant the Cat was up a tree, hiding among the leaves.
Egyptian	لا هوه عايز ياكل أكلنا ولا سايبنا ناكله!
Levantine	م رضي ياكل من الاكل الي احنا عملناه وفوق كده كمان م تركنا ناكله!
MSA	لا يريد أن يأكل طعامنا ولا يريدنا أن نأكل طعامنا.
English	He doesn’t want to eat our food and yet he will not let us eat it!
Egyptian	السلحفة كان فرحان أوي.
Levantine	السلحفاة كان كتير مبسوط.
MSA	كان السلحف سعيدا للغاية.
English	The Tortoise was very glad.
Egyptian	كان بيحب يروح المسرح ويلعب بالكوتشينة ويستمتع بالرياضات زي سباق الخيل.
Levantine	كان بيحب يروح ع المسرح ويلعب الورق وبيحب الرياضة زي سباق الخيل.
MSA	كان يحب الذهاب إلى المسرح ولعب الورق والاستمتاع بالرياضات مثل سباق الخيل.
English	He liked to go to the theater, play cards and enjoy sports such as horse racing.
Egyptian	المرأة بقى ليها حق التصويت سنة ١٩١٥. وجرى انتخاب أول ست في البرلمان سنة ١٩٢٢.
Levantine	النسوان انعطوا حق التصويت في ١٩١٥. أول عضوة إنثى في البرلمان تم انتخابها في ١٩٢٢.
MSA	منحت المرأة حق التصويت عام ١٩١٥. وقد تم انتخاب أول امرأة في البرلمان عام ١٩٢٢.
English	Women were given the right to vote in 1915. The first female member of parliament was elected in 1922.

Table 8: Translation examples from the *test* set