Investigating Code-Mixed Modern Standard Arabic-Egyptian to English Machine Translation

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

Abstract

Recent progress in neural machine translation (NMT) has made it possible to translate successfully between monolingual language pairs where large parallel data exist, with pre-trained models improving performance even further. Although there exists work on translating in code-mixed settings (where one of the pairs includes text from two or more languages), it is still unclear what recent success in NMT and language modeling exactly means for translating code-mixed text. We investigate one such context, namely MT from code-mixed Modern Standard Arabic and Egyptian Arabic (MSAEA) into English. We develop models under different conditions, employing both (i) standard end-to-end sequence-to-sequence (S2S) Transformers trained from scratch and (ii) pre-trained S2S language models (LMs). We are able to acquire reasonable performance using only MSA-EN parallel data with S2S models trained from scratch. We also find LMs fine-tuned on data from various Arabic dialects to help the MSAEA-EN task. Our work is in the context of the Shared Task on Machine Translation in Code-Switching. Our best model achieves **25.72** BLEU, placing us first on the official shared task evaluation for MSAEA-EN.

1 Introduction

Recent year have witnessed fast progress in various areas of natural language processing (NLP), including machine translation (MT) where neural approaches have helped boost performance when translating between pairs with especially large amounts of parallel data. However, tasks involving a need to process data from different languages mixed together remain challenging for all NLP tasks. This phenomenon of using two or more languages simultaneously in speech or text is referred to as *code-mixing* (Gumperz, 1982) and is prevalent in multilingual societies (Sitaram et al., 2019). Code-mixing is challenging since the space of possibilities when processing mixed data is vast, but also because there is not usually sufficient code-mixed resources to train models on. Nor is it clear how much code-mixing existing language models may have seen during pre-training, and so ability of these language models to transfer knowledge to downstream code-mixing tasks remain largely unexplored.

In this work, we investigate translation under a code-mixing scenario where sequences at source side are a combination of two varieties of the collection of languages referred to as Arabic. More

	(1) MSAEA	أنا عايز شغل جامد يا جدعان.
English	**Human**	I want hard work, guys.
	GMT	I want a rigid job, Jadaan.
	S2ST	I want a solid job, jadan.
	(2) MSAEA	الدكاترة قالو انى مش همشي طبيعى تانى.
English	**Human**	The doctors said I can't walk normally again.
	GMT	The doctors said that I was not a normal marginal again.
	S2ST	Doctors said I wasn't a natural marginality again.

Table 1: Code-mixed Modern Standard Arabic-Egyptian Arabic (MSAEA) sentences with their English human translation, Google machine translation (GMT)[1], and translation by a sequence-to-sequence Transformer model (S2ST) trained from scratch on 55M MSA-English parallel sentences. **Green** refers to good translations. **Red** refers to erroneous translation.

[1]We use Google Translate API https://cloud.google.com/translate.

Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, pages 56–64

https://doi.org/10.26615/978-954-452-056-4_008

specifically, we take as our objective translating between Modern Standard Arabic (MSA) mixed with Egyptian Arabic (EA) (source; collectively abbreviated here as MSAEA) into English (target). Table1 shows two examples of MSAEA sentences and their human and machine translations. We highlight problematic translations caused by nixing of Egyptian Arabic with MSA. Through work related to the shared task, we target the following three main research questions:

1. *How do models trained from scratch on purely MSA data fare on the code-mixed MSAEA data (i.e., the zero-shot EA setting)?*

2. *How do existing language models perform under the code-mixed condition (i.e., MSAEA)?*

3. *What impact, if any, does exploiting dialectal Arabic (DA) data (i.e., from a range of dialects) have on the MSAEA code-mixed MT context?*

Our main contributions in this work lie primarily in answering these three questions. We also develop powerful models for translating from MSAEA to English.

The rest of the paper is organized as follows: Section 2 discusses related work. The shared task is described in Section 3. Section 4 describes external parallel data we exploit to build our models. Section 5 presents the proposed MT models. Section 6 presents our experiments, and our different settings. We provide evaluation on Dev data in Section 7 and official results in Section 8. We conclude in Section 9.

2 Related Work

A thread of research on code-mixed MT focuses on automatically generating synthetic code-mixed data to improve the downstream task. This includes attempts to generate linguistically-motivated sequences (Pratapa et al., 2018). Some work leverages sequence-to-sequence (S2S) models (Winata et al., 2019) to generate code-mixing exploiting an external neural MT system, while others (Garg et al., 2018) use a recurrent neural network along with data generated by a sequence generative adversarial network (SeqGAN) and grammatical information such as from a part of speech tagger to generate code-mixed sequences. These methods have dependencies and can be costly to scale beyond one language pair.

Arabic MT. For Arabic, some work has focused on translating between MSA and Arabic dialects. For instance, Zbib et al. (2012) studied the impact of combined dialectal and MSA data on dialect/MSA to English MT performance. Sajjad et al. (2013) uses MSA as a pivot language for translating Arabic dialects into English. Salloum et al. (2014) investigate the effect of sentence-level dialect identification and several linguistic features for MSA/dialect-English translation. Guellil et al. (2017) propose an neural machine translation (NMT) system for Arabic dialects using a vanilla recurrent neural networks (RNN) encoder-decoder model for translating Algerian Arabic written in a mixture of Arabizi and Arabic characters into MSA. Baniata et al. (2018) present an NMT system to translate Levantine (Jordanian, Syrian, and Palestinian) and Maghrebi (Algerian, Moroccan, Tunisia) to MSA, and MSA to English. Farhan et al. (2020), propose unsupervised dialectal NMT, where the source dialect is not represented in training data. This last problem is referred to as zero-shot MT (Lample et al., 2018).

DA Arabic MT Resources. There are also efforts to develop dialectal Arabic MT resources. For example, Meftouh et al. (2015) present the Parallel Arabic Dialect Corpus (PADIC),[2] which is a multi-dialect corpus including MSA, Algerian, Tunisian, Palestinian, and Syrian. Recently, Sajjad et al. (2020a) also introduced AraBench, an evaluation suite for dialectal Arabic to English MT. AraBench consists of five publicly available datasets: Arabic-Dialect/English Parallel Text (APT) (Zbib et al., 2012), Multi-dialectal Parallel Corpus of Arabic (MDC) (Bouamor et al., 2014), MADAR Corpus (Bouamor et al., 2018), Qatari-English speech corpus (Elmahdy et al., 2014), and the English Bible translated into MSA.[3]

3 Code-Switching Shared Task

The goal of the shared tasks on machine translation in code-switching settings[4] is to encourage building MT systems that translate a source sentence into a target sentence while one of the directions contains

[2] https://sites.google.com/site/torjmanepnr/6-corpus

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

[4] https://code-switching.github.io/2021.

an alternation between two languages (i.e., code-switching). We note that, in the current paper, we employ the wider term *code-mixing*. The shared task involves two subtasks:

1. **Supervised MT.** For supervised MT, gold data are provided to participants for training and evaluating models that take English as input and generate Hinglish sequences.

2. **Unsupervised MT.** In this subtask, the goal is to develop systems that can generate high quality translations for multiple language combinations. These combinations include Spanish-English to English or Spanish, English to Spanish-English, Modern Standard Arabic-Egyptian Arabic (MSAEA) to English and vice versa. For each pair, only test data are provided to participants, with no reference translations.

In the current work, we focus on the unsupervised MT subtask only. More specifically, we build models exclusively for MSAEA to English. Our approach exploits external data to train a variety of models. We now describe these external datasets.

4 Parallel Datasets

4.1 MSA-English Data

In order to develop Arabic MT models that can translate efficiently across different text domains, we make use of a large collection of parallel sentences extracted from the Open Parallel Corpus (OPUS) (Tiedemann, 2012). OPUS contains more than 2.7 billion parallel sentences in 90 languages. To train our models, we extract more than $\sim$ 61M sentences MSA-English parallel sentences from the whole collection. Since OPUS can have noise and duplicate data, we clean this collection and remove duplicates before we use it. We now describe our quality assurance method for cleaning and deduplication of the data.

Data Quality Assurance. To keep only high quality parallel sentences, we follow two steps:

1. We run a cross-lingual semantic similarity model (Yang et al., 2019) on each pair of sentences, keeping only sentences with a bilingual similarity score between 0.30 and 0.99. This allows us to filter out sentence pairs whose source and target are identical (i.e., similarity score = 1) and those that are not good translations of one another (i.e., those with a cross-lingual semantic similarity score < 0.3).

2. Observing some English sentences in the source data, we perform an analysis based on sub-string matching between source and target, using the word trigram sliding window method proposed by Barrón-Cedeño and Rosso (2009) and used in Abdul-Mageed et al. (2021) to de-duplicate the data splits. In other words, we compare each sentence in the source side (i.e., MSA) to the target sentence (i.e., English). We then inspect all pairs of sentences that match higher than a given threshold, considering thresholds between 90% and 30%. We find that a threshold of $> 75\%$ safely guarantees completely distinct source and target pairs.

More details about the MSA-English OPUS dataset before and after our quality assurance, including deduplication, are provided in Table 2.

Data	**#Sentences**
Bible	62.2K
EUbookshop	1.7K
GlobalVoices	52.6K
Gnome	150
Infopankki	50.8K
KDE4	116.2K
MultiUN	9.8M
News Commentary	90.1K
OpenSubtitles	29.8M
QED	500.9K
Tanzil	187K
Tatoeba	27.3K
TED2013	152.8K
Ubuntu	6K
UN	74.1K
UNPC	20M
Wikipedia	151.1K
Total	61M
Similarity $\in$ **[**0.3 **-** 0.99**]**	5.7M
***N*-gram deduplication (>**0.75**)**	55.2M

Table 2: Parallel datasets extracted from OPUS (Tiedemann, 2012). We remove duplicate and identical pairs, keeping only high quality translations.

Dataset	Egyptian	Levantine	Gulf
Bouamor et al. (2018)	18K	22K	26K
Elmahdy et al. (2014)	–	–	14.7K
Zbib et al. (2012)	38K	138K	–
Total	56K	160K	40.7K

Table 3: Our parallel DA-English datasets. Gulf comprises data from Bouamor et al. (2018), Elmahdy et al. (2014) , and Zbib et al. (2012).

4.2 Dialectal Arabic-English Data

Several recent works show that MT models trained on one dialect can be used to improve models targeting other dialects (Farhan et al., 2020; Sajjad et al., 2020b). For this reason, we exploit several parallel dialectal Arabic (DA)-English datasets in order to enhance the MSAEA to English translation.

DA-English Parallel Corpus. Zbib et al. (2012) provide 38k Egyptian Arabic (EA)-English and 138k Levantine-English sentences ($\sim$ 3.5 million tokens of Arabic dialects), collected from online user groups and dialectal Arabic weblogs. The authors use crowdsourcing to translate this dataset into English.

MADAR Corpus. MADAR Bouamor et al. (2018) is a commissioned dataset where 26 Arabic native speakers were tasked to translate 2k English sentences each into their own native dialect. In addition, Bouamor et al. (2018) translate 10k more sentences for five selected cities: Beirut, Cairo, Doha, Cairo, Tunis, and Rabat. The MADAR dataset also has region-level categorization (i.e., Gulf, Levantine, Nile, and Maghrebi). In our work, we use only the Gulf, Levantine, and Nile (Egyptian) dialects, and exclude Maghrebi.[5]

Qatari-English Speech Corpus. This parallel corpus comprises 14.7k Qatari-English sentences collected by Elmahdy et al. (2014) from talk-show programs and Qatari TV series.

More details about all our parallel dialectal-English datasets are in Table 3.

4.3 Data Splits and Pre-Processing

Data Splits. For our experiments, we split the MSA and DA data as follows:

MSA. We randomly pick 10k sentences for validation (MSA-Dev) from MSA parallel data (see Section 4.1) after cleaning, and we use the rest of this data ($\sim$ 55.14M) for training (MSA-Train).

DA. For validation (DA-Dev), we randomly pick 6k sentences from the 38k Egyptian-English data provided by Zbib et al. (2012). We then use the rest of the data (i.e., $\sim$ 250.7k) for training (DA-Train).

Pre-Processing. Pre-processing is an important step for building any MT model as it can significantly affect end results (Oudah et al., 2019). For all our models, we only perform light pre-processing in order to retain a faithful representation of the original (naturally occurring) text. We remove diacritics and replace URLs, user mentions, and hashtags with the generic string tokens `URL`, `USER`, and `HASHTAG` respectively. Our second step for pre-processing is specific to each type of models we train as we will explain in the respective sections.

5 MT Models

5.1 From-Scratch Seq2Seq Models

We train our models on the MSA-English parallel data described in section 4.1 on MSA-Train with a Transformer (Vaswani et al., 2017) model as implemented in Fairseq (Ott et al., 2019). For that, we follow Ott et al. (2018) in using 6 blocks for each of the encoder and decoder parts. We use a learning rate of 0.25, a dropout of 0.3, and a batch size $4,000$ tokens. For the optimizer, we use Adam (Kingma and Ba, 2014) with beta coefficients of 0.9 and 0.99 which control an exponential decay rate of running averages, with a weight decay of 10^{-4}. We also apply an inverse square-root learning rate scheduler with a value of $5e^{-4}$ and $4,000$ warm-up updates. For the loss function, we use label smoothed cross entropy with a smoothing strength of 0.1. We run the Moses tokenizer (Koehn et al., 2007) on our input before passing data to the model. For vocabulary, we use a joint Byte-Pair Encoding (BPE) (Sennrich et al., 2015) vocabulary with 64K split operations for subword segmentation.

5.2 Pre-Trained Seq2Seq Language Models

We also fine-tune two state-of-the-art pre-trained multlingual generative models, mT5 (Xue et al., 2020) and mBART (Liu et al., 2020) on DA-Train for 100 epochs. We use early stopping during fine-tuning and identify the best model on DA-Dev. We use the HuggingFace (Wolf et al., 2020) implemen-

[5]We do not make use of the Maghrebi data due to the considerable linguistic differences between Maghrebi and the the Egyptian dialect we target in this work.

Source:	مش عارفين نتأكد و مش عارفين البنات فين.
S2ST	we don't know for sure and the girls don't know finn .
mT5	we can't make sure and we don't know where the girls are
mBART	we don't know where to make sure and we don't know where the girls are
Source:	انا عايز اعرف موقف الاخوان الرسمي من التحرش بالحريري و نجاد البرعي ولو البلطجية دول مش تبعهم الرئيس يستعمل سلطته.
S2ST	i want to know the brothers' official position on harassment of liberals and nejad al-barai, even the thugs, countries that are not followed by the president are using his authority and ordering their immediate arrest.
mT5	i want to know the situation of the official brothers from harassment of the silky and najad albarea and if these pants are not their president the president uses his power and order to arrest them immediately
mBART	i want to know the position of the official brothers from harassment in the army and najad al-bara'y, even if these are not theirs , the president should use his authority and order to arrest them immediately
Source:	" user : عاوزين محامي يروح معانا القسم يا جدعان صبحي ووليد تليفوناتهم مقفولة."
S2ST	user: there is a need for a lawyer to help the section, jadan sobhi and walid televonas closed .
mT5	user: we want a lawyer to go with us to the section , guys , sobhe and waleed their telephones are closed
mBART	« user : we want a lawyer to go with us to the section, oh good morning, and their telephones are closed. »
Source	بيعقدوا الجلسات في أماكن مش محاكم ، و ما ينفعش خلق الله يدخلوا من غير تصريح ، عشان المتهمين لما ييجوا يمنعوهم و يحكموا غيابي !!
S2ST	they hold hearings in places where there are no courts, and what thrives on god's creation will enter without permission, because the accused will not prevent them and judge my absence!
mT5	they have sessions in places that are not courts, and god doesn't allow people to enter without a permit, so that when they come and prevent them and rule me absence
mBART	they hold meetings in places where there is no courts, and god doesn't allow people to enter without a permit, so that when the accused come they stop them and rule them

Table 4: MSA-EA sentences with their English translations using our Models. **S2ST:** Sequence-to-sequence Transformer model trained from scratch. Data samples are extracted from the shared task Test data. **Green** refers to good translation. **Red** refers to problematic translation.

tation of each of these models, with the default settings for all hyper-parameters.

6 Experiments and Settings

In this section, we describe the different ways we fine-tune and evaluate our models.

6.1 Zero-Shot Setting

First, we use S2ST model trained on MSA-English data exclusively to evaluate MSAEA code-mixed data . While we can refer to this setting as *zero-shot*, we note that it is not truly zero-shot in the strict sense of the word due to the code-mixed nature of the data (i.e., the data has a mixture of MSA and EA). Hence, we will refer to this setting as *zero-shot EA*.

6.2 Fine-Tuning Setting

Second, we further fine-tune the three models (i.e., S2ST, mT5, and mBART) on the DA data described in Section 4.2. While the downstream shared task data only involves EA mixed with MSA, we follow Farhan et al. (2020) and Sajjad et al. (2020b) in fine-tuning on different dialects when targeting a single downstream dialect (EA in our case). We will simply refer to this second setting as *Fine-Tuned DA*.

Model	Setting	Blue
S2ST	Zero Shot EA	8.54
	Fine-tuned DA	9.33
	Zero Shot EA (true-cased)	11.59
	Fine-tuned DA (true-cased)	12.57
mT5	Fine-tuned DA	24.70
	Fine-tuned DA (true-cased)	26.35
mBART	Fine-tuned DA	23.80
	Fine-tuned DA (true-cased)	**26.07**

Table 5: Results of models on DA-Dev data. **S2ST:** Sequence-to-sequence Transformer model trained from scratch. We note that in the zero-shot EA setting the S2ST model is trained on 55M bitext sentences.

7 Evaluation on Dev Data

We report results of all our models under different settings in BLEU scores (Papineni et al., 2002). In addition to evaluation on uncased data, we run a language modeling based truecaser (Lita et al., 2003) on the outputs of our different models.[6] Results presented in Table 5 show that S2ST achieves relatively low scores (between 8.54 and 12.57) on all settings. In comparison, both mBART and mT5 fine-tuned on DA-Train are able to translate MSAEA to English with BLEU scores of 23.80 and 24.70 respectively. We note that truecasing the output results in improving the results with an average of +2.55 BLEU points.

8 Official Shared Task (Test) Results

Table 7 shows results of all our MT models with different settings on the official shared task Test set. We observe that the Transformer model in the *zero shot EA* setting (a model that does not see Egyptian Arabic data) was able to translate MSAEA to English with 21.34 BLEU. As expected, fine-tuning all the models on DA-Train improves results across all models and leads to the best BLEU score of 25.72% with the S2ST model.

Comparing performance of the S2ST model on Dev and Test data, we observe that Test data results are better. This suggests that Test data comprises more MSA than EA sequences. To test this hypothesis, we run a binary MSA-DA classifier Abdul-Mageed et al. (2020) on both the Dev and Test data to acquire MSA and DA distributions on each dataset. Results of this analysis, shown in Table 6, confirm our hypothesis about Test data involving significantly more MSA (i.e., 72.31%) compared to Dev data.

[6]We were not been able to report results based on truecasing in this paper, but we note that we will provide these results in the camera ready version of this paper.

Dataset	#Size	MSA	DA
DA-Dev	6,164	18.36%	81.64%
Official Test	6,500	72.31%	27.69%

Table 6: The data distribution (MSA Vs DA) in the DA-Dev and the official Test set.

Model	Setting	Blue
S2ST	Zero Shot EA	21.34
	Fine-tuned DA	22.51
	Zero Shot EA (true-cased)	23.68
	Fine-tuned DA (true-cased)	**25.72**
mT5	Fine-tuned DA	16.41
	Fine-tuned DA (true-cased)	18.80
mBART	Fine-tuned DA	17.17
	Fine-tuned DA (true-cased)	19.79

Table 7: Results of our models on official Test data. Again, in the zero-shot EA setting the S2ST model is trained on 55M bitext sentences,

Discussion. We inspect output translations from our models on Test data. We observe that even though S2ST performs better than the two language models on Test data, both of these models are especially able to translate Egyptian Arabic tokens such as فين in example (1) in Table 4 well. Again, Test data contain more MSA than DA as we explained earlier and hence the S2ST model (which is trained on 55M sentence pairs) outperforms each of the two language models. This analysis suggests that fine-tuning the language models on more MSA-ENG should result in better performance.

Returning to our three main research questions, we can reach a number of conclusions. For **RQ1**, we observe that models trained from scratch on purely MSA data fare reasonably well on the code-mixed MSAEA data (i.e., *zero-shot EA* setting). This is due to lexical overlap between MSA and EA. For **RQ2**, we also note that language models such as mT5 and mBART do well under the code-mixed condition, more so than models trained from scratch when inference data involve more EA. This is the case even though these language models in our experiments are fine-tuned with significantly

less data (i.e., $\sim$ 250K pairs) than the from-scratch S2ST models (which are trained on 55M MSA + 250K DA pairs). For **RQ3**, our results show that training on data from various Arabic dialects helps translation in the MSAEA code-mixed condition. This is in line with previous research (Farhan et al., 2020) showing that exploiting data from various dialects can help downstream translation on a single dialect dialect in the zero-shot setting.

9 Conclusion

We described our contribution to the shared tasks on MT in code-switching.[7] Our models target the MSAEA to English task under the unsupervised condition. Our experiments show that training models on MSA data is useful for the MSAEA-to-English task in the zero-shot EA setting. We also show the utility of pre-trained language models such as mT5 and mBART on the code-mixing task. Our models place first in the official shared task evaluation. In the future, we intend to apply our methods on other dialects of Arabic and investigate other methods such as backtranslation for improving overall performance.

Acknowledgements

We gratefully acknowledges support from the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, Canadian Foundation for Innovation, Compute Canada (www.computecanada.ca) and UBC ARC-Sockeye (https://doi.org/10.14288/SOCKEYE) and Penguin Computing POD™ (pod.penguincomputing.com).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *arXiv preprint arXiv:2101.01785*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Dinesh Pabbi, Kunal Verma, Rannie Lin, et al. 2021. Mega-cov: A billion-scale dataset of 100+ languages for covid-19. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3402–3420.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018.

Alberto Barrón-Cedeño and Paolo Rosso. 2009. On automatic plagiarism detection based on n-grams comparison. In *European conference on information retrieval*, pages 696–700. Springer.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245.

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 *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a tv broadcasts speech recognition system for qatari arabic. In *LREC*, pages 3057–3061.

Wael Farhan, Bashar Talafha, Analle Abuammar, Ruba Jaikat, Mahmoud Al-Ayyoub, Ahmad Bisher Tarakji, and Anas Toma. 2020. Unsupervised dialectal neural machine translation. *Information Processing & Management*, 57(3):102181.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018. Code-switched language models using dual RNNs and same-source pretraining. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3083, Brussels, Belgium. Association for Computational Linguistics.

Imane Guellil, Faical Azouaou, and Mourad Abbas. 2017. Neural vs statistical translation of algerian arabic dialect written with arabizi and arabic letter. In *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC*, volume 31, page 2017.

John J. Gumperz. 1982. *Discourse Strategies*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. 2007. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In *Final Report of the Johns Hopkins 2006 Summer Workshop*.

[7] https://code-switching.github.io/2021

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. Truecasing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 152–159.

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

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.

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

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*.

Mai Oudah, Amjad Almahairi, and Nizar Habash. 2019. The impact of preprocessing on arabic-english statistical and neural machine translation. *arXiv preprint arXiv:1906.11751*.

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.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020a. Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020b. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on 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.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. 2019. A survey of code-switched speech and language processing. *CoRR*, abs/1904.00784.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. 2012:2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

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

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch.

2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.