Data Augmentation Techniques for Machine Translation of Code-Switched Texts: A Comparative Study

Injy Hamed,[1,2] **Nizar Habash,**[1] **Ngoc Thang Vu**[2]
[1]Computational Approaches to Modeling Language Lab, New York University Abu Dhabi
[2]Institute for Natural Language Processing, University of Stuttgart
{injy.hamed,nizar.habash}@nyu.edu
thang.vu@ims.uni-stuttgart.de

Abstract

Code-switching (CSW) text generation has been receiving increasing attention as a solution to address data scarcity. In light of this growing interest, we need more comprehensive studies comparing different augmentation approaches. In this work, we compare three popular approaches: lexical replacements, linguistic theories, and back-translation (BT), in the context of Egyptian Arabic-English CSW. We assess the effectiveness of the approaches on machine translation and the quality of augmentations through human evaluation. We show that BT and CSW predictive-based lexical replacement, being trained on CSW parallel data, perform best on both tasks. Linguistic theories and random lexical replacement prove to be effective in the lack of CSW parallel data, where both approaches achieve similar results.

1 Introduction

Code-switching (CSW) is the alternation of language in text or speech, which can occur across different levels of granularity: sentences, words and morphemes. CSW is a common phenomenon in Arabic-speaking countries, as in other multilingual communities. Given that Arabic is a morphologically rich language (Habash et al., 2012), speakers produce morphological CSW, as illustrated below:

حالاً algorithm+الـ implement+هـ طيب ⇐

'*Okay*, *I'll*+implement *the*+algorithm *right away*'

CSW introduces a set of challenges to NLP systems, not least of which is data scarcity. This is attributed to CSW being a predominantly spoken phenomenon, only recently increasing in written form on social media. Data augmentation has proved to be a successful workaround for this limitation. Researchers have investigated several techniques for CSW data augmentation, including learning CSW points (Solorio and Liu, 2008; Gupta et al., 2021), lexical replacements (Appicharla et al., 2021; Xu and Yvon, 2021; Gupta et al., 2021; Hamed et al., 2022c), linguistic theories (Pratapa et al., 2018; Lee et al., 2019; Hussein et al., 2023), neural-based approaches (Chang et al., 2018; Winata et al., 2018, 2019; Menacer et al., 2019; Song et al., 2019; Li and Vu, 2020), and machine translation (MT) (Vu et al., 2012; Tarunesh et al., 2021). With increasing efforts in this area, we need more comparative studies to better understand the merits and requirements of different approaches.

Efforts along these lines include the work of Pratapa and Choudhury (2021), where different linguistic-driven and lexical replacement techniques were compared through human evaluation, but not for NLP tasks. Winata et al. (2018) propose the use of pointer-generator network and compare it against the equivalence constraint (EC) theory (Poplack, 1980) and random lexical replacement for LM, without human evaluation. Hamed et al. (2022c) compare multiple lexical replacement techniques covering human evaluation and performance on language modeling (LM), automatic speech recognition (ASR), MT, and speech translation. Hussein et al. (2023) compare using the EC theory and random lexical replacement for LM and ASR, also reporting human assessments.

In this work, we compare three main approaches: **lexical replacements**, **linguistic theories**, and **back-translation (BT)**. We evaluate the approaches for both naturalness of CSW generations and performance on MT, where we focus on CSW Egyptian Arabic-English to English translation. The rationale for our focus on MT is the scarcity of work around data augmentation as opposed to LM and ASR. Furthermore, previous work on MT focuses on lexical replacements (Menacer et al., 2019; Song et al., 2019; Appicharla et al., 2021; Xu and Yvon, 2021; Gupta et al., 2021; Hamed et al., 2022c) and BT (Tarunesh et al., 2021), without substantial comparison between approaches. Through our comparative study, we provide answers to the following research questions:

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 140–154
December 6-10, 2023

- **RQ1:** Which augmentation technique perform best in zero-shot and non-zero-shot settings (with/without the availability of CSW parallel corpora) for MT?
- **RQ2:** Does generating more natural synthetic CSW sentences entail improvements in MT?

2 Data Augmentation Techniques

We provide an overview on the investigated techniques.[1] Our aim is to augment Arabic-to-English parallel sentences, converting the source side of the parallel data from monolingual Arabic to CSW Arabic-English, further extending the MT training data with CSW instances. In Figure 1, we provide an example showing possible augmentations across techniques. More examples are shown in Table 4.

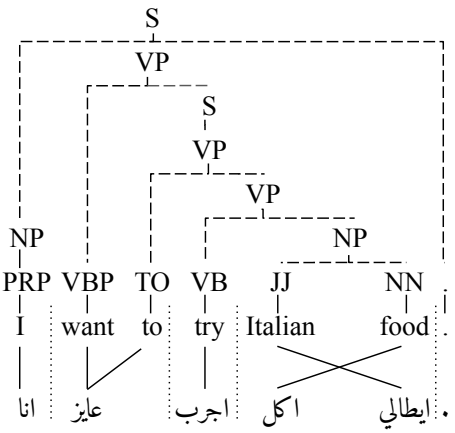
2.1 Lexical Replacements

We investigate the following three approaches:

Dictionary Replacement: We replace x random Arabic words on the source side with English gloss entries. We obtain the gloss entries using MADAMIRA (Pasha et al., 2014), an Arabic morphological analyzer and tagger. Such a specialized analysis system is required for this task as Arabic is morphologically rich and orthographically ambiguous. We refer to this approach as LEX$_{Dict}$.

Aligned with Random CSW Point Assignment: We augment the Arabic-to-English parallel sentences by randomly picking x source-target aligned words (using intersection alignments) and replacing the source words with their counterpart words on the target side. In Hamed et al. (2022c), the authors investigated two types of alignments for performing source-target replacements: (1) word replacements using intersection alignments and (2) segment replacements where grow-diag-final alignments are used to identify aligned segments. Given that segment replacements were shown to be superior, we follow that setup in our experiments. We refer to this approach as LEX$_{Rand}$.

Aligned with Learnt CSW Point Prediction: Similar to the previous approach, we perform target-to-source replacements; however, the choice of words on the target side to be inserted into the source side is based on a CSW predictive model (Appicharla et al., 2021; Hamed et al., 2022c). The model is trained to identify words on the target side that would be plausible CSW words on the source side. The task of CSW point prediction is modeled as a sequence-to-sequence classification task. The neural network takes as input the target sentence word sequence $x = \{x_1, x_2, .., x_N\}$, where N is the length of the sentence. The network outputs a sequence $y = \{y_1, y_2, .., y_N\}$, where $y_n \in \{1,0\}$ represents whether the word x_n is a plausible CSW word or not. To obtain the training data for the predictive model, we utilize a limited amount of CSW Egyptian Arabic-English to English parallel sentences, where we tag the words on the target side as 0 or 1 based on whether they appear as CSW words on the source side or not. This is done using a matching algorithm described in Hamed et al. (2022c). The CSW predictive model is then trained by fine-tuning mBERT on this data.[2] Afterwards, to augment Arabic-to-English parallel data, we use the model to identify CSW candidates on the target side which are inserted in the source side using segment replacements. For a detailed description of this approach, see Hamed et al. (2022c). We refer to this approach as LEX$_{Pred}$.



Approach	Augmentation Example
LEX$_{Dict}$	انا ⇐ wanting اجرب اكل ايطالي .
LEX$_{Rand}$	انا ⇐ want to اجرب اكل ايطالي .
LEX$_{Pred}$	انا عايز اجرب اكل ⇐ Italian .
EC	انا عايز ⇐ try اكل ايطالي .
EC & ML	انا عايز اجرب ⇐ Italian food .

Figure 1: An example showing possible augmentations by the different techniques. We show the parse tree for the English sentence and word alignments. The permissible switching points under the EC theory are shown by the dotted lines.

[1] We make our relevant code available at: http://arzen.camel-lab.com/

[2] The hyperparameters are shown in Appendix C.

2.2 Linguistic Theories

We cover the following two linguistic theories:

Equivalence Constraint (EC) Theory: The EC Theory (Poplack, 1980) is an alternational model for CSW, where there are no defined matrix and embedded languages. Instead, the theory states that code-switching can occur at points where the surface structures of both languages map onto each other. In the example in Figure 1, the permissible alternations are indicated by dotted lines. Generating "أكل Italian" and "Italian أكل" is not allowed as the syntactic rules of both languages are different (Arabic adjectives follow the nouns they modify).

Matrix Language Frame (MLF) Theory: The MLF Theory (Myers-Scotton, 1997), on the other hand, is an insertional model. It is based on the identification of a matrix language, to which constituents of the embedded language are inserted such that the sentence follows the grammatical structure of the matrix language, and the embedded language is inserted at grammatically correct points. Unlike the EC theory, replacements from the embedded to matrix language are not allowed within nesting sub-trees. Replacements of closed-class constituents are also not allowed, including determiners, quantifiers, prepositions, possessive, auxiliaries, tense morphemes, and helping verbs.

For both linguistic theories, we use the GCM tool (Rizvi et al., 2021).[3] The tool provides multiple augmentations per source-target parallel sentence, following a linguistic theory. To sample from these generations, it provides two sampling approaches: random and Switch-Point Fraction (SPF) (Pratapa et al., 2018). In random sampling, k generations are picked randomly. In SPF sampling, the generations are ranked based on their SPF distribution compared to a reference distribution obtained from real CSW data and the top-k generations are chosen. SPF is calculated as the number of switch points divided by the total number of language-dependent tokens in a sentence.[4] We set k to 1, which is unified across all techniques. We include both sampling approaches, where we refer to the variants as EC$_{Rand}$, EC$_{SPF}$, ML$_{Rand}$, and ML$_{SPF}$.

[3]The tool takes $\approx$ 12 hours to augment 309k parallel sentences for each linguistic theory.

[4]The current version of the GCM tool provides an implementation for Switch Point (SP) which does not account for the number of tokens in the sentence. We implement our own code for ranking based on SPF. See footnote 1.

	Model	top-k	#Aug
1	[en-csw&ar]	1	0.1k
2	[en-csw&ar]→[en-csw]	1	10k
3	[en-csw&ar]+[en-en]→[en-csw]	1	19k
4	[en-csw&ar]+[en-en]→[en-csw]	19	151k

Table 1: The number of CSW generations (#Aug) obtained from the different BT setups: (1) BT model trained on English to Arabic and English to CSW Arabic-English parallel sentences, (2) same as 1 and followed by fine-tuning using the English to CSW Arabic-English parallel data, (3) similar to 2, with appending English sentences to both sides of the training data, and (4) same as 3 with utilizing the top-19 hypotheses.

2.3 Back-translation

Despite BT (Sennrich et al., 2015) being a well-known data augmentation technique, it has received little attention in the scope of CSW (Tarunesh et al., 2021). In this approach, we train a BT model to translate English sentences to CSW Arabic-English. We then use this model to translate the target side of the Arabic-to-English parallel sentences, generating synthetic CSW Arabic-English to English parallel sentences. The BT model is trained on a limited amount of English to CSW Arabic-English parallel sentences and a larger amount of English to Arabic parallel data. However, when using this model to translate 309k English sentences, only 109 CSW sentences are generated, with the rest of the translations being monolingual Arabic. This is due to the training data of the BT model only constituting of 0.7% of sentences having CSW. We boost the number of generated CSW synthetic sentences through the following steps:

1. We fine-tune the model using the English to CSW Arabic-English parallel data.
2. In the BT model training data, we further append the English sentences in the parallel corpus to both source and target sides.
3. At inference, instead of obtaining the top-1 hypothesis for each English sentences, we utilize the top-k hypotheses and obtain the CSW translation with the highest confidence score. We set k to 19, where we could not further increase the value of k due to computational constraints.

In Table 1, we show the effect of each step on the number of obtained CSW generations, reaching a total of 151k CSW augmentations by applying all three steps (augmenting 49% of original sentences).

3 Experimental Setup

3.1 Data

We use two sources of data: (1) ArzEn-ST (Hamed et al., 2022b), which is a CSW-focused parallel corpus and (2) monolingual Egyptian Arabic-to-English parallel corpora. ArzEn-ST contains English translations of a CSW Egyptian Arabic-English speech corpus (Hamed et al., 2020) gathered through informal interviews with bilingual speakers. The corpus is divided into train, dev, and test sets having $3.3k$, $1.4k$, and $1.4k$ sentences (containing $2.2k$, $0.9k$, and $0.9k$ CSW sentences).

For Egyptian Arabic-to-English parallel sentences, we obtain $309k$ parallel sentences from the following parallel corpora: Callhome Egyptian Arabic-English Speech Translation Corpus (Gadalla et al., 1997; LDC, 2002b,a; Kumar et al., 2014), LDC2012T09 (Zbib et al., 2012), LDC2017T07 (Chen et al., 2017), LDC2019T01 (Chen et al., 2019), LDC2021T15 (Tracey et al., 2021), and MADAR (Bouamor et al., 2018). The corpora cover web (LDC2012T09/LDC2019T01), chat (LDC2017T07/LDC2021T15), and conversational (Callhome/MADAR) domains. We use the corpora data splits if pre-defined, otherwise, we follow the guidelines provided by Diab et al. (2013). Data preprocessing is discussed in Appendix B.

3.2 Setup of Augmentation Approaches

Through augmentation, we convert the source side of the $309k$ Arabic-to-English parallel sentences to CSW Arabic-English. For word alignments, we use Giza++ (Och and Ney, 2003).[5] For the augmentation approaches that require CSW parallel sentences, we utilize ArzEn-ST. In BT, we train the model on the train sets of the parallel corpora outlined in Section 3.1, with reversed source and target sides. The predictive model in LEX$_{Pred}$ is trained on the portion of ArzEn-ST train set having CSW sentences. That subset is also utilized in the linguistic theories to obtain the reference SPF distribution ($= 0.22$). It is also utilized in LEX$_{Dict}$ and LEX$_{Rand}$, where the value of x is set to 19% of the source words based on the percentage of English words in ArzEn-ST train set CSW sentences, which is 18.8%. However, the average percentage calculated over sentences is 22.1% with a standard deviation of 17.5%. The decision of 19% is in agreement with Hussein et al. (2023) where the authors report LM perplexities achieved by embedding different percentages of English words in Arabic text using random lexical replacement and decide on a percentage of 20%. In future work, we believe an interesting direction is to model CSW distribution to obtain a wider coverage of various CSW levels rather than targeting a single percentage for all sentences.

3.3 Machine Translation System

We train a Transformer model using Fairseq (Ott et al., 2019) on a single GeForce RTX 3090 GPU. We use the hyperparameters from the FLORES benchmark for low-resource machine translation (Guzmán et al., 2019).[6] The hyperparameters are given in Appendix C. We use a BPE model trained jointly on source and target sides with a vocabulary size of $16k$ (which outperforms $1, 3, 5, 8, 32, 64k$). The BPE model is trained using Fairseq with character_coverage set to 1.0. For MT training data, we use the train sets of the corpora outlined in Section 3.1. For the augmentation experiments, we append the synthetically generated CSW Arabic-English to English parallel sentences. For development and evaluation of the MT models, we use ArzEn-ST dev and test sets.

4 Evaluation

In this section, we present intrinsic evaluation, human evaluation, and extrinsic evaluation.

4.1 Intrinsic Evaluation

In Table 2, we report the number of CSW sentences generated per technique as well as CSW statistics. We report that the number of augmentations varies considerably across techniques: LEX$_{Dict}$ > LEX$_{Rand}$ > BT > EC > LEX$_{Pred}$ > ML.

With regards to CSW metrics, we report Code-mixing Index (CMI) (Gambäck and Das, 2016), SPF, and the average percentage of English tokens over sentences. CMI reflects the level of mixing between multiple languages, and is calculated on the sentence-level as follows:

$$CMI(x) = \frac{\frac{1}{2}*(N(x) - max_{L_i \in \mathbf{L}}\{t_{L_i}\}(x)) + \frac{1}{2}P(x)}{N(x)}$$

[5] Following Hamed et al. (2022c), in lexical replacements, we take the union of grow-diag-final alignments trained on word and stem spaces. We use the same alignments in linguistic theories, as it produces more generations compared to the default alignment setup used in the GCM tool (grow-diag-final-and alignments trained on word space using fast-align (Dyer et al., 2013)).

[6] FLORES hyperparameters outperformed Vaswani et al. (2017) in Gaser et al. (2022) on the same utilized datasets.

	Size (*k*)	CMI	SPF	SPF$_\sigma$	%En
ArzEn-ST	-	0.21	0.22	0.13	22.1
LEX$_{Dict}$	239.6	0.28	0.33	0.12	22.5
LEX$_{Rand}$	192.7	0.25	0.24	0.12	31.9
LEX$_{Pred}$	112.9	0.24	0.22	0.13	36.8
EC$_{Rand}$	142.1	0.30	0.29	0.14	59.0
EC$_{SPF}$	142.1	0.25	0.24	0.08	64.4
ML$_{Rand}$	98.2	0.27	0.27	0.14	60.8
ML$_{SPF}$	98.2	0.25	0.25	0.10	63.1
BT	151.1	0.18	0.19	0.14	65.2

Table 2: The number of generated sentences per technique, and their CMI and SPF mean and standard deviation (SPF/SPF$_\sigma$) values and average percentage of English words (%En). We also report the figures for the CSW sentences in ArzEn-ST train set as reference.

	Understandability
1	No, this sentence doesn't make sense.
2	Not sure, but I can guess the meaning of this sentence.
3	Certainly, I get the meaning of this sentence.
	Naturalness
1	Unnatural, and I can't imagine people using this style of code-mixed Arabic-English.
2	Weird, but who knows, it could be some style of code-mixed Arabic-English.
3	Quite natural, but I think this style of code-mixed Arabic-English is rare.
4	Natural, and I think this style of code-mixed Arabic-English is used in real life.
5	Perfectly natural, and I think this style of code-mixed Arabic-English is very frequently used.

Table 3: The evaluation dimensions for human evaluation, following Pratapa and Choudhury (2021).

where N is the number of language-dependent tokens in sentence x; $L_i \in \mathbf{L}$ is the set of languages in the corpus; $max_{L_i \in \mathbf{L}}\{t_{L_i}\}$ is the number of tokens in the dominating language in x; and P is the number of switch points in x, where $0 \leq P < N$. The corpus-level CMI is calculated as the average of sentence-level CMI values.

We observe that in general, LEX$_{Rand}$ and LEX$_{Pred}$ provide the closest figures to ArzEn-ST with regards to CSW metrics. It is to be noted that unlike LEX$_{Rand}$ and SPF-based linguistic theories, no explicit CSW heuristics were provided to LEX$_{Pred}$, and the predictive model learnt to imitate the CSW frequency in ArzEn-ST. In the case of linguistic theories, we note that SPF sampling provides CMI and SPF figures that are closer to ArzEn-ST than random sampling. Finally, we report that the linguistic theories and BT augmentations contain high percentages of English words.

4.2 Human Evaluation

In order to assess the quality of the synthetically generated CSW sentences, we perform a human evaluation study. Out of the original sentences that get augmented by all techniques, we randomly sample 150 sentences.[7] These sentences are evaluated by three annotators across the eight augmentation techniques against two measures: understandability and naturalness. All three annotators are female Egyptian Arabic-English bilingual speakers, in the age range of 33-39, all graduates of private English schools. We follow the rubrics introduced by Pratapa and Choudhury (2021), outlined in Table 3. Understandability is rated on a scale of 1-3 and naturalness is rated on a scale of 1-5 where scores of 3-5 are assigned to natural sentences with different levels of commonality to be encountered in real life. A total of 1,200 augmentations are annotated by each of the three annotators for both understandability and naturalness, giving a total of 7,200 annotations.[8] For each augmentation, we calculate the mean opinion score (MOS) as the average of scores received by the three annotators. The full results are provided in Appendix E, where the percentage of sentences falling under each MOS range per technique is presented in Table 7. In Figure 2, we show the percentage of sentences perceived as natural by annotators across techniques (summation of the last two rows in Table 7). We observe the following ranking between techniques: BT > LEX$_{Pred}$ > ML > EC > LEX$_{Rand}$ > LEX$_{Dict}$.

With regards to linguistic theories, as noted by Doğruöz et al. (2021), computational implementations of linguistic theories do not necessarily generate natural CSW sentences that would mimic human CSW generation. We elaborate on this point in Section 6. While ML achieves higher naturalness ratings than EC, we do not observe superiority across the different sampling techniques, which can be due to the SPF values only changing slightly between both techniques in our case.

[7] The sentences are sampled uniformly across the six corpora used in data augmentation to have equal representation of the different domains (web/chat/conversational).

[8] The annotation task took an average of 9 hours per annotator, and each annotator was paid $160.

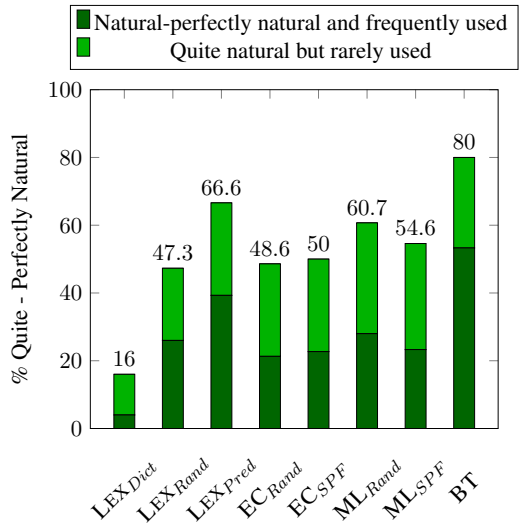


Figure 2: The percentage of augmentations with 3$\leq$ MOS $\leq$5 (quite natural but rarely used - perfectly natural and frequently used) per technique.

This can be different in other setups with different reference SPF distributions. With regards to understandability, there is less variability across the techniques (91-96% of the augmentations are given ratings between 2 and 3), except for LEX$_{Dict}$ (the percentage is 65%). We perform inter-annotator agreement by applying pairwise Cohen Kappa (Cohen, 1960), reporting 0.25-0.28 (fair agreement) on naturalness between annotator pairs. Low agreement on this task is expected, as CSW attitude is speaker-dependent (Vu et al., 2013). The pairwise Cohen Kappa scores for understandability are higher (0.33-0.35), yet still showing fair agreement. We also apply Fleiss' Kappa (Fleiss, 1971) across all annotators, scoring fair agreement of 0.312 and 0.249 for understandability and naturalness.[9]

4.3 Extrinsic MT Evaluation

The augmentation techniques covered in this study vary in terms of requirements. One main difference is the reliance on CSW parallel data, which is only available for a few CSW language pairs (Hamed et al., 2022b). To have a fair comparison and to show the effectiveness of the techniques in both cases (availability and lack of CSW-focused parallel corpora), we run two sets of experiments:

- Zero-shot setting: In this setting, our baseline system is trained only using the 309k monolingual Arabic-to-English parallel sentences. We extend the training data with augmentations generated using techniques that do not require CSW parallel data, namely: LEX$_{Dict}$, LEX$_{Rand}$, EC, and ML.
- Non-zero-shot setting: In this setting, we assume the availability of CSW parallel data. We train our baseline system using the monolingual Arabic-to-English parallel sentences in addition to ArzEn-ST corpus. We then append the augmentations generated by each of the investigated techniques.

In the following sections, we present our baseline systems and the results for zero-shot and non-zero-shot settings. The full results are reported in Table 5, showing BLEU (Papineni et al., 2002), chrF, chrF++ (Popović, 2017), and BERTScore (F1) (Zhang et al., 2019). BLEU, chrF and chrF++ are calculated using SacreBLEU (Post, 2018). We report performance on ArzEn-ST test set; on all sentences as well as CSW sentences only. Our analysis in this section is based on chrF++. This choice is based on chrF++ showing higher correlation with human judgments over chrF (Popović, 2017) and chrF showing higher correlation over BLEU (Kocmi et al., 2021). We report performance on ArzEn-ST test set CSW sentences, as this is our main concern. Statistical significance tests for zero- and non-zero-shot settings are shown in Table 6.

4.3.1 Baselines

We develop the following MT baselines, showing the improvements achieved by each source of data:

- BL$_{CSW}$: We train it solely on ArzEn-ST train set, having 3.3k parallel sentences.
- BL$_{Mono}$: We train it on the 309k monolingual Arabic-to-English parallel sentences.
- BL$_{MonoTgt}$: In BL$_{Mono}$, we observe that English words on the source side get dropped in translation. This issue has been previously tackled by researchers using techniques including direct copying (Song et al., 2019) or the use of a pointer network (Menacer et al., 2019). We propose a simple technique of including target-target pairs in the training process. In other words, in addition to the source-target sentences used in BL$_{Mono}$, we append the English (target) sentences on both source and target sides, ending up with 617k parallel sentences. Our hypothesis is that by doing so, the model learns to retain the English words on the source side through translation.

[9] We use the implementation provided in: https://github.com/Shamya/FleissKappa/blob/master

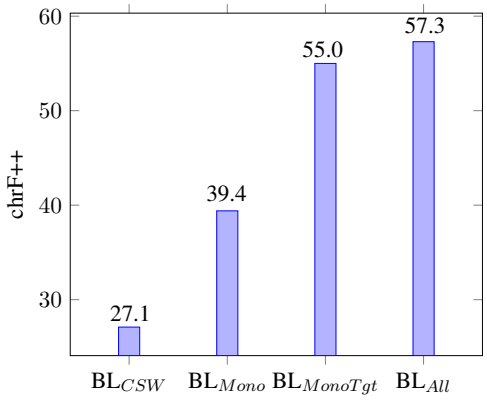


Figure 3: chrF++ scores of the different baselines on ArzEn-ST test set CSW sentences.

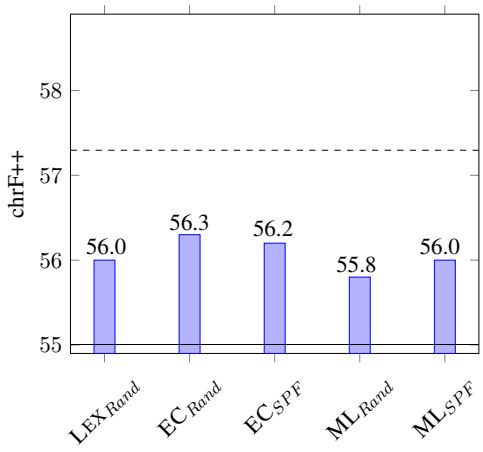


Figure 4: The effectiveness of the augmentation techniques in a zero-shot setting. We show the chrF++ scores on ArzEn-ST test set CSW sentences. The solid and dashed lines represent BL$_{MonoTgt}$ and BL$_{All}$.

- BL$_{All}$: We include the same data as in BL$_{MonoTgt}$, in addition to ArzEn-ST train set, giving a total of 620k parallel sentences.

The chrF++ scores are shown in Figure 3 (full results in Table 5 Exp 1-4). The effectiveness of the simple step of adding target-target pairs during training is confirmed, where BL$_{MonoTgt}$ achieves an increase of +15.6 chrF++ points over BL$_{Mono}$. Adding ArzEn-ST train set (BL$_{All}$) results in further +2.3 chrF++ points, achieving 57.3 on chrF++.

4.3.2 Zero-shot Setting Experiments

This setting is tailored to the majority of CSW language pairs, that are under-resourced and lack CSW-focused parallel corpora. We demonstrate the effectiveness of the augmentation techniques in a zero-shot setting. Given that LEX$_{Pred}$ and BT are reliant on CSW parallel data, they are excluded from this comparison. We include the following approaches: LEX$_{Dict}$, LEX$_{Rand}$, EC$_{Rand}$, EC$_{SPF}$, ML$_{Rand}$, and ML$_{SPF}$. We acknowledge that some of these approaches rely on heuristics obtained from CSW data, such as SPF value or the enforced CSW percentage. However, we argue that these figures can be obtained from textual data (that is more easily accessible than parallel data). The baseline in this setting is BL$_{MonoTgt}$, which is our best baseline that does not utilize real CSW data.

We report that LEX$_{Dict}$ degrades the MT performance, falling 3.2 chrF++ points below the baseline. We present the chrF++ scores for the other techniques in Figure 4 (full results in Table 5 Exp 5-10). We observe that linguistic-based models and LEX$_{Rand}$ perform equally well, despite LEX$_{Rand}$ generating more data. As shown in Table 6, there is no statistical significance between LEX$_{Rand}$ and linguistic-based models. Comparing the linguistic theories, EC performs better than ML, however, there is no difference between SPF and random sampling strategies. Overall, EC$_{Rand}$ performs the best, with statistical significance over ML$_{Rand}$ and ML$_{SPF}$, achieving +1.3 chrF++ points over BL$_{MonoTgt}$.

4.3.3 Non-zero-shot Experiments

In this setting, we assume the availability of CSW-focused parallel data, and thus compare all augmentation techniques. The baseline for this setting is BL$_{All}$. The chrF++ scores are shown in Figure 5 (full results in Table 5 Exp 11-18).[10]

LEX$_{Dict}$ falls below BL$_{All}$ by 1.4 chrF++ points, we thus exclude it from Figure 5. We observe that LEX$_{Pred}$ and BT outperform LEX$_{Rand}$ and linguistic theories. The best performance is achieved by BT, achieving +1.3 chrF++ points over BL$_{All}$. We also report that LEX$_{Rand}$ and linguistic theories are unable to achieve significant improvements over BL$_{All}$.[11] We examine the amount of real in-domain CSW data that would result in equivalent performance achieved by LEX$_{Rand}$ and linguistic theories in the zero-shot setting. In Figure 6, we show a learning curve by adding different amounts of ArzEn-ST train set CSW sentences to BL$_{MonoTgt}$ training data, and show that LEX$_{Rand}$ and linguis-

[10]The lexical replacements results differ from Hamed et al. (2022c) due to the following design decisions: (1) we append target-target pairs in the training data; and (2) we only include generated CSW sentences, and not sentences that get fully converted to English, in order to better control for variables.

[11]The improvement achieved by ML$_{Rand}$ over BL$_{All}$ is not statistically significant.

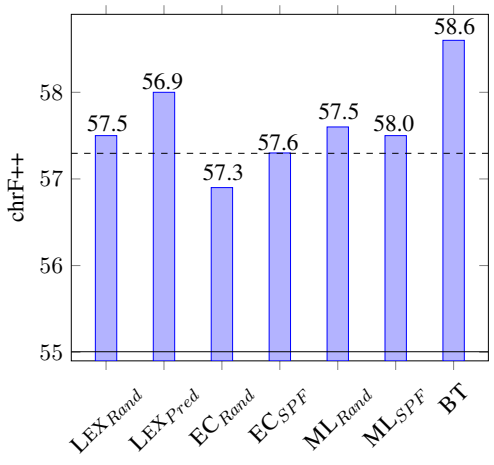


Figure 5: The effectiveness of the augmentation techniques in a non-zero-shot setting. We show the chrF++ scores on ArzEn-ST test set CSW sentences. The solid and dashed lines represent BL$_{MonoTgt}$ and BL$_{All}$.

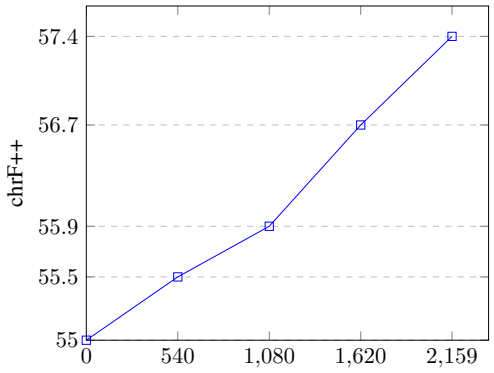


Figure 6: Learning curve of adding different amounts of ArzEn-ST train set CSW sentences to the BL$_{MonoTgt}$ training data. We show the chrF++ scores on ArzEn-ST test set CSW sentences.

tic theories (generating 98-192k CSW synthetic sentences) perform on par at 50% of ArzEn-ST train set CSW sentences ($\approx$ 1,080 sentences).

5 Discussion

In this section, we revisit our RQs:

RQ1 - Which augmentation techniques perform the best for MT? In the zero-shot setting, LEX$_{Rand}$ and linguistic theories achieve similar performance, with EC outperforming ML models. In the non-zero shot setting, BT outperforms all techniques, followed by LEX$_{Pred}$. Both techniques, being trained on real CSW data, are able to generate more natural CSW sentences, that could also be closer in CSW style to ArzEn-ST.

RQ2 - Does generating more natural synthetic CSW sentences entail improvements in MT? Here, we look into the relation between MT scores and naturalness ratings. In the non-zero shot setting, we report a correlation of 0.97 between the chrF++ scores (presented in Figure 5) and the percentage of sentences perceived as natural (presented in Figure 2). This demonstrates a strong positive correlation between MT performance and naturalness of augmentations.

Given that the number of augmentations varies considerably across techniques (shown in Table 2), this variation could empower some techniques over others, affecting performance as an effect of quantity rather than quality. Therefore, we perform another set of experiments where we control for this variable. We report results under a constrained setup, where we restrict the augmentations appended to the baseline training data to only those that are successfully augmented across all techniques (= 24.8k sentences). We first append the constrained augmentations per technique to BL$_{MonoTgt}$ training data. The results are presented in Table 5 Exp 19-26. The order based on chrF++ is: BT > LEX$_{Pred}$ > [LEX$_{Rand}$ & linguistic theories] > LEX$_{Dict}$. The correlation between the chrF++ scores achieved on ArzEn-ST test set CSW sentences and the percentage of sentences perceived as natural is 0.95. We replicate the constrained experiments with appending the constrained augmentations to BL$_{All}$ training data. Given the availability of CSW data in the training data, and with constraining the amounts of augmented data, the majority of the models show no improvements over BL$_{All}$. We therefore cannot use this setup to make conclusions on the relation between quality and performance. However, from the previously discussed findings, we confirm a positive relation between the naturalness of generated synthetic sentences and MT performance.

6 Insights into Augmentations

In this section, we present insights into the augmentations produced by the different techniques, further elaborating on their strengths and weaknesses. All examples mentioned in this section refer to the examples demonstrated in Table 4.

Lexical Replacements: The main drawback in LEX$_{Dict}$ is that the replaced words might not be correct translations within context, which can negatively affect the MT model. As shown in Table 4

Example 1, اولع *AwlE*[12] 'turn on', in the context of *turn on this light*, is replaced by 'kindle'. This drawback is also observed in the case of ambiguous words, as shown in Example 2, where the word طابع *TAbE* 'stamp' is replaced by 'impression'. With regards to LEX_{Rand}, CSW can occur at unnatural locations, such as replacing ده *dh* 'this' in Example 1. This is less likely for LEX_{Pred}, which is reflected in human evaluation.

Linguistic theories: We observe that applying linguistic theories does not guarantee naturalness, e.g., the augmentation provided by EC_{SPF} shown in Example 3, despite being a correct augmentation under the EC theory, was given a rating of '2' by all annotators. Moreover, the effectiveness of these techniques is tied to (and currently restricted by) the performance of the available tools that implement them. We observe that the augmentations obtained from the GCM tool in some cases violate the EC or ML theories. For the EC theory, in Example 4, we demonstrate a case where Arabic-to-English alternation occurs at the word 'station' which is a point of syntactic divergence in محطة الاوتوبيس *mHTp AlAwtwbys* and 'the coach station'. For the ML theory, in Example 5, the augmentation includes the stand-alone CSW segment 'in', while replacements of closed-class constituents (including prepositions) is prohibited. As the tool relies on generating Arabic parse trees from English parse trees using alignments, errors are likely to be introduced. Furthermore, as noted by Hussein et al. (2023), the augmentations are sometimes missing information from the original sentences.[13]

BT: We observe that BT is capable of generating correct morphological code-switching (MCS). As shown in Example 6, the MCS construction 'بت+handle' *bt+handle* 'handles' is correctly composed of بت *bt* 'progressive-imperfect-2nd-masc-sing' preceding the verb 'handle' and 'field' is correctly preceded by the definite article ال *Al* 'the'. While researchers have provided insights into common Arabic-English MCS constructs (Kniaź, 2017; Kniaź and Zawrotna, 2021; Hamed et al., 2022a), there is no current research that allows for modeling Arabic-English MCS in a rule-based approach. Therefore, the ability of neural-based approaches to generate MCS is an advantage. On the other hand, similar to the partial transcription issue noted in Chowdhury et al. (2021) for ASR models using BPE, the BT approach can provide partial translations of words, such as 'modifications' translated to s+تعديل *tEdyl+s* 'modification+s'(Example 7). BT might also provide literal translation. With both issues combined, we find cases such as 'locker' being translated to er+قفل *qfl+er* 'lock+er'.

[12]Buckwalter Arabic Transliteration (Habash et al., 2007).

[13]We reduce the effect of this issue by validating that the generated sentences are complete using the alternational matrices computed by the tool in the generation process, and giving priority to sampling from validated augmentations.

7 Conclusion and Future Work

We present a comparative study between different CSW data augmentation techniques and their effectiveness for MT in both zero-shot and non-zero-shot settings. We show that in the zero-shot setting, random lexical replacement performs equally well as linguistic theories. In the case of non-zero shot setting, back-translation performs best, followed by CSW predictive-based lexical replacement. Both approaches also stand out in human evaluation, where we confirm a positive correlation between naturalness of augmentations and MT performance. However, both approaches are reliant on expensive and limited CSW parallel data. Overall, the set of approaches examined proves useful in alleviating data scarcity. Each approach comes with particular merits and requirements, guiding the choice for different research needs. In future work, we plan on enhancing the back-translation approach to leverage larger amounts of English data. In parallel, we will investigate the effectiveness of generative AI to broaden the benchmark of approaches, and expand our study to cover other NLP tasks.

Limitations

One limitation of the presented work is that the models were evaluated on one test set only, and therefore, we cannot interpret how the models will perform on other sets, covering other domains and sources (spoken versus written). Another limitation is that the study involves only one language pair. Further research is needed to investigate whether the findings hold for other language pairs. A third limitation is the low variability in the annotators' demographics, as the three annotators are female annotators, in the same age group, receiving similar levels of education. Including a broader set of annotators would enrich research with insights on the level of agreement between annotators with wider background differences.

Ethics Statement

We could not identify any ethical issues in the work, and to our best knowledge, we believe it complies with the ACL Ethics Policy. We use ArzEn-ST corpus, which is distributed under an Attribution-ShareAlike 4.0 International license, where we adhere to its intended usage. All other parallel corpora are also publicly available, including MADAR, Callhome, and LDC corpora.

References

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. IITP-MT at CALCS2021: English to Hinglish neural machine translation using unsupervised synthetic code-mixed parallel corpus. In *Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching (CALCS)*, pages 31–35.

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 LREC*, pages 3387–3396.

Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee. 2018. Code-switching sentence generation by generative adversarial networks and its application to data augmentation. In *Proceedings of Interspeech*, pages 554–558.

Song Chen, Dana Fore, Stephanie Strassel, Haejoong Lee, and Jonathan Wright. 2017. BOLT Egyptian Arabic SMS/Chat and Transliteration LDC2017T07. Philadelphia: Linguistic Data Consortium.

Song Chen, Jennifer Tracey, Christopher Walker, and Stephanie Strassel. 2019. BOLT Arabic discussion forum parallel training data LDC2019T01. Philadelphia: Linguistic Data Consortium.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic ASR. In *Proceedings of Interspeech*, pages 2466–2470.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Mona Diab, Nizar Habash, Owen Rambow, and Ryan Roth. 2013. LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.

A Seza Doğruöz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *In Proceedings of ACL-IJCNLP*, pages 1654–1666.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of NAACL-HLT*, pages 644–648.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic transcripts LDC97T19. Web Download. Philadelphia: Linguistic Data Consortium.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of LREC*, pages 1850–1855.

Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. Exploring segmentation approaches for neural machine translation of code-switched Egyptian Arabic-English text. In *Proceedings of EACL*, pages 86–100.

Abhirut Gupta, Aditya Vavre, and Sunita Sarawagi. 2021. Training data augmentation for code-mixed translation. In *Proceedings of NAACL-HLT*, pages 5760–5766.

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 EMNLP-IJCNLP*, pages 6098–6111.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A morphological analyzer for Egyptian Arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology*, pages 1–9.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22.

Injy Hamed, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2022a. Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech. *Computer Speech & Language*, 72:101278.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022b. ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *Proceedings of WANLP*, pages 119–130.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022c. Investigating lexical replacements for Arabic-English code-switched data augmentation. In *Proceedings of LoResMT Workshop*, pages 86–100.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched

Egyptian Arabic-English. In *Proceedings of LREC*, pages 4237–4246.

Amir Hussein, Shammur Absar Chowdhury, Ahmed Abdelali, Najim Dehak, Ahmed Ali, and Sanjeev Khudanpur. 2023. Textual data augmentation for Arabic-English code-switching speech recognition. In *Proceedings of SLT*, pages 777–784.

Małgorzata Kniaź. 2017. English lexical items in Egyptian Arabic. Code-switching or borrowing? *Alicante Journal of English Studies*, 30:187–212.

Małgorzata Kniaź and Magdalena Zawrotna. 2021. Embedded english verbs in Arabic-English code-switching in Egypt. *International Journal of Bilingualism*, 25(3):622–639.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Conference on Machine Translation*, pages 478–494.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177–180.

Gaurav Kumar, Yuan Cao, Ryan Cotterell, Chris Callison-Burch, Daniel Povey, and Sanjeev Khudanpur. 2014. Translations of the CALLHOME Egyptian Arabic corpus for conversational speech translation. In *Proceedings of IWSLT*, pages 244–248.

LDC. 2002a. 1997 HUB5 Arabic transcripts – LDC2002T39. Web Download. Philadelphia: Linguistic Data Consortium.

LDC. 2002b. CALLHOME Egyptian Arabic transcripts supplement LDC2002T38. Web Download. Philadelphia: LDC.

Grandee Lee, Xianghu Yue, and Haizhou Li. 2019. Linguistically motivated parallel data augmentation for code-switch language modeling. In *Proceedings of Interspeech*, pages 3730–3734.

Chia-Yu Li and Ngoc Thang Vu. 2020. Improving code-switching language modeling with artificially generated texts using cycle-consistent adversarial networks. In *Proceedings of Interspeech*, pages 1057–1061.

Mohamed Amine Menacer, David Langlois, Denis Jouvet, Dominique Fohr, Odile Mella, and Kamel Smaïli. 2019. Machine translation on a parallel code-switched corpus. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 426–432.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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 (Demonstrations)*, pages 48–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of LREC*, pages 1094–1101.

Shana Poplack. 1980. Sometimes i'll start a sentence in Spanish y termino en Español: Toward a typology of code-switching. *The bilingualism reader*, 18(2):221–256.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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

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 ACL*, pages 1543–1553.

Adithya Pratapa and Monojit Choudhury. 2021. Comparing grammatical theories of code-mixing. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 158–167.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of EACL (System Demonstrations)*, pages 205–211.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*, pages 86–96.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for Arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of WANLP*, pages 167–177.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of EMNLP*, pages 973–981.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of NAACL-HLT, Volume 1 (Long and Short Papers)*, pages 449–459.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. In *Proceedings of ACL-IJCNLP*, pages 3154–3169.

Jennifer Tracey et al. 2021. BOLT Egyptian Arabic SMS/chat parallel training data LDC2021T15. Philadelphia: Linguistic Data Consortium.

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 *Proccedings of NeurIPS*, pages 5998–6008.

Ngoc Thang Vu, Heike Adel, and Tanja Schultz. 2013. An investigation of code-switching attitude dependent language modeling. In *Proceedings of the International Conference on Statistical Language and Speech Processing*, pages 297–308.

Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li. 2012. A first speech recognition system for Mandarin-English code-switch conversational speech. In *Proceedings of ICASSP*, pages 4889–4892.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Learn to code-switch: Data augmentation using copy mechanism on language modeling. *arXiv preprint arXiv:1810.10254*.

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 CoNLL*, pages 271–280.

Jitao Xu and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching (CALCS)*, pages 84–94.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris CallisonBurch. 2012. Machine translation of Arabic dialects. In *Proceedings of NAACL*, pages 49–59.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. In *Proceedings of ICLR*.

A Augmentation Examples

In Table 4, we present examples of augmentations generated by the different techniques. These examples are discussed in Section 6.

B Data Preprocessing

Following Hamed et al. (2022c), we remove corpus-specific annotations, remove URLs and emoticons through tweet-preprocessor, tokenize numbers, apply lowercasing, run Moses' (Koehn et al., 2007) tokenizer as well as MADAMIRA (Pasha et al., 2014) simple tokenization (D0), and perform Alef/Ya normalization.[14] For entries with words having literal and intended translations, we opt for one translation having all literal translations and another having all intended translations. For LDC2017T07, we utilize the work by Shazal et al. (2020), where the authors used a sequence-to-sequence model to transliterate the corpus text from Arabizi (where Arabic words are written in Roman script) to Arabic orthography. For the Egyptian Arabic-to-English parallel corpora discussed in Section 3.1, we only utilize the 309k monolingual Egyptian Arabic-to-English parallel sentences available in these corpora, where we do not utilize the parallel sentences with code-switching within the scope of this work. In future work, it would be interesting to investigate how the effectiveness of data augmentation varies with the availability of different amounts of real CSW parallel data, to draw further conclusions under different levels of low-resourcefulness. Also, for MADAR and LDC2012T09, we only utilize the Egyptian Arabic subsets of both corpora.

C Hyperparameters

For finetuning mBERT for the CSW predictive model, we set the epochs to 5, drop-out rate to 0.1, warmup steps to 500, batch size to 13, and learning rate to 0.0001. The training and inference time took $\approx$ 12 hours.

For MT, we use the following train command: python3 fairseq_cli/train.py $DATA_DIR –source-lang src –target-lang tgt –arch transformer –share-all-embeddings –encoder-layers 5 –decoder-layers 5 –encoder-embed-dim 512 –decoder-embed-dim 512 –encoder-ffn-embed-dim 2048 –decoder-ffn-embed-dim 2048 –encoder-attention-heads 2 –decoder-attention-heads 2 –encoder-normalize-before –decoder-normalize-before –dropout 0.4 –attention-dropout 0.2 –relu-dropout 0.2 –weight-decay 0.0001 –label-smoothing 0.2 –criterion label_smoothed_cross_entropy –optimizer adam –adam-betas '(0.9, 0.98)' –clip-norm 0 –lr-scheduler inverse_sqrt –warmup-updates 4000 –warmup-init-lr 1e-7 –lr 1e-3 –stop-min-lr 1e-9 –max-tokens 4000 –update-freq 4 –max-epoch 100 –save-interval 10 –ddp-backend=no_c10d

D MT Results

In Table 5, we report the MT results, showing BLEU, chrF, chrF++, and BERTScore(F1). The

[14] https://pypi.org/project/tweet-preprocessor/

Examples		
1	Src	ازاي اولع النور ده ؟
	Tgt	how can i turn on this light ?
	LEX$_{Dict}$	ازاي kindle النور ده ؟
	LEX$_{Rand}$	ازاي اولع النور this ؟
	LEX$_{Pred}$	ازاي اولع light ده ؟
	EC$_{Rand}$	how can i turn on النور ده ؟
	EC$_{SPF}$	ازاي اولع this light ؟
	ML$_{Rand}$	ازاي اولع light ده ؟
	ML$_{SPF}$	how can i turn on this النور ؟
	BT	ازاي اقدر افتح ال light ده ؟
2	Src	جبت طابع تاني ؟
	Tgt	got another stamp ?
	LEX$_{Dict}$	جبت impression تاني ؟
3	Src	لازم ناخد جواب من الادارة في طنطا كل شهرين
	Tgt	we must take a letter from the management in tanta ; every two months
	EC$_{Rand}$	لازم ناخد جواب من الادارة in tanta ; every شهرين
4	Src	هو محطة الاوتوبيس فين ؟
	Tgt	where 's the coach station ?
	EC$_{Rand}$	هو station الاوتوبيس فين ؟
5	Src	انا سبت محفظتي في محلك .
	Tgt	i left my wallet in your shop .
	ML$_{Rand}$	انا left محفظتي in محلك .
6	Src	لو انت مبتتعاملش مع المجال ده ، ممكن تقترح محل متخصص في ده ؟
	Tgt	if you don 't handle that field , could you suggest a store specializing in it ?
	BT	لو انت مش بت handle ال field ده ، ممكن تقترح محل متخصص فيه ؟
7	Src	فيه اي تعديلات انت عايز تعملها ؟
	Tgt	are there any modifications you would like to make ?
	BT	في اي تعديلs تحب تعملها ؟

Table 4: Examples of synthetic CSW sentences generated by the different augmentation techniques, demonstrating strengths and weaknesses of techniques. Given that Arabic is written from right to left, we display all augmentations in a right-to-left orientation.

statistical significance between the models in the zero-shot and non-zero-shot settings for chrF++ achieved on ArzEn-ST test set CSW sentences are shown in Table 6. The number of parameters in the models for Exp 1 is 39,712,768 and Exp 2-26 is 44,967,936. The training time taken by Exp 1 is $\approx$ 8 minutes, Exp 2 $\approx$ 2.6 hours, and Exp 3-26 $\approx$ 5.2-6.5 hours.

E Human Evaluation

We present the full results of the human evaluation study discussed in Section 4.2. For each evaluated augmentation, we calculate the mean opinion score (MOS) as the average of scores received by the three annotators. In Table 7, we present the percentage of sentences falling under each MOS range for understandability and naturalness per augmentation technique. In Table 8, we present the average MOS scores per technique.

			All Test Sentences				CSW Test Sentences			
Exp	**Model**	**|Train|**	**BLEU**	**chrF**	**chrF++**	**BertScore(F1)**	**BLEU**	**chrF**	**chrF++**	**BertScore(F1)**
Baselines										
1	BL$_{CSW}$	3,340	8.3	27.2	26.6	0.218	8.3	27.8	27.1	0.175
2	BL$_{Mono}$	308,689	22.2	42.1	41.4	0.387	20.7	39.9	39.4	0.315
3	BL$_{MonoTgt}$	617,378	31.7	54.9	53.5	0.519	32.8	56.5	55.0	0.510
4	BL$_{All}$	620,718	**34.4**	**57.4**	**55.7**	**0.547**	**35.6**	**59.1**	**57.3**	**0.549**
Zero-shot Experiments										
5	+LEX$_{Dict}$	857,022	29.8	52.3	51.1	0.499	30.2	53.1	51.8	0.474
6	+LEX$_{Rand}$	810,030	32.8	55.7	54.3	**0.531**	34.1	57.5	56.0	**0.529**
7	+EC$_{Rand}$	759,478	**33.7**	**56.1**	**54.7**	0.528	**34.9**	**57.8**	**56.3**	0.522
8	+EC$_{SPF}$	759,478	33.1	55.8	**54.5**	0.530	34.5	**57.6**	**56.2**	0.527
9	+ML$_{Rand}$	715,610	32.6	55.5	54.2	0.527	33.9	57.2	55.8	0.520
10	+ML$_{SPF}$	715,610	33.0	55.8	54.4	0.529	34.2	57.4	56.0	0.523
Non-zero-shot Experiments										
11	+LEX$_{Dict}$	860,362	33.6	56.0	54.5	0.536	34.8	57.6	55.9	0.530
12	+LEX$_{Rand}$	813,370	34.2	57.1	55.5	0.546	35.9	59.2	57.5	0.546
13	+LEX$_{Pred}$	733,660	35.2	57.5	56.1	**0.550**	36.8	59.5	58.0	0.551
14	+EC$_{Rand}$	762,818	33.5	56.6	55.1	0.544	34.9	58.6	56.9	0.547
15	+EC$_{SPF}$	762,818	34.6	57.0	55.5	0.547	36.2	59.0	57.3	0.549
16	+ML$_{Rand}$	718,950	34.9	57.4	55.8	0.548	36.3	59.3	57.6	0.549
17	+ML$_{SPF}$	718,950	34.3	57.3	55.7	0.548	35.7	59.2	57.5	0.547
18	+BT	771,793	**35.8**	**58.2**	**56.6**	**0.550**	**37.5**	**60.3**	**58.6**	**0.553**
Constrained Experiments										
19	+LEX$_{Dict}$	642,221	30.4	53.3	52.0	0.502	31.2	54.5	53.0	0.482
20	+LEX$_{Rand}$	642,221	32.2	55.3	53.9	0.529	33.3	56.9	55.5	0.524
21	+LEX$_{Pred}$	642,221	32.9	55.8	54.3	0.530	34.3	57.6	56.1	0.527
22	+EC$_{Rand}$	642,221	32.2	55.5	54.0	0.525	33.6	57.4	55.7	0.521
23	+EC$_{SPF}$	642,221	32.7	55.3	53.9	0.526	34.1	57.2	55.6	0.520
24	+ML$_{Rand}$	642,221	32.3	55.3	53.9	0.524	33.4	56.9	55.4	0.517
25	+ML$_{SPF}$	642,221	32.5	55.3	53.9	0.523	33.9	57.0	55.5	0.518
26	+BT	642,221	**34.3**	**56.4**	**55.0**	**0.534**	**36.1**	**58.4**	**56.9**	**0.531**

Table 5: We report the MT results (BLEU, chrF, chrF++, and BertScore) on ArzEn-ST test set, for all sentences as well as CSW sentences only. We report the results of the baselines (Section 4.3.1), zero-shot (Section 4.3.2), non-zero-shot (Section 4.3.3), and constrained (Section 5) settings. The best performing models in each setting are bolded. The overall best performing model is underlined.

		LEX_{Dict}	LEX_{Rand}	EC_{Rand}	EC_{SPF}	ML_{Rand}
	chrF++	51.8	56.0	56.3	56.2	55.8
LEX_{Dict}	51.8					
LEX_{Rand}	56.0	0.0010*				
EC_{Rand}	56.3	0.0010*	0.0539			
EC_{SPF}	56.2	0.0010*	0.1139	0.2208		
ML_{Rand}	55.8	0.0010*	0.0959	0.0030*	0.0120*	
ML_{SPF}	56.0	0.0010*	0.2667	0.0240*	0.0649	0.1518

(a) Statistical significance between the models in the zero-shot setting.

		LEX_{Dict}	LEX_{Rand}	LEX_{Pred}	EC_{Rand}	EC_{SPF}	ML_{Rand}	ML_{SPF}
	chrF++	55.9	57.5	58.0	56.9	57.3	57.6	57.5
LEX_{Dict}	55.9							
LEX_{Rand}	57.5	0.0010*						
LEX_{Pred}	58.0	0.0010*	0.0190*					
EC_{Rand}	56.9	0.0010*	0.0040*	0.0010*				
EC_{SPF}	57.3	0.0010*	0.1359	0.0030*	0.0280*			
ML_{Rand}	57.6	0.0010*	0.2957	0.0310*	0.0020*	0.1149		
ML_{SPF}	57.5	0.0010*	0.4096	0.0240*	0.0030*	0.1239	0.3137	
BT	58.6	0.0010*	0.0010*	0.0040*	0.0010*	0.0010*	0.0010*	0.0010*

(b) Statistical significance between the models in the non-zero-shot setting.

Table 6: Statistical significance between models in the zero- and non-zero-shot settings calculated on the chrF++ scores achieved on ArzEn-ST test set CSW sentences. We present the p-values and mark p-values < 0.05 with $*$, where the null hypothesis can be rejected. We include the chrF++ scores for easier readability and comparison.

MOS	LEX_{Dict}	LEX_{Rand}	LEX_{Pred}	EC_{rand}	EC_{spf}	ML_{rand}	ML_{spf}	BT
Understandability								
$1 \leq * < 2$	35.3	4.0	4.0	7.3	8.0	8.7	9.3	6.0
$2 \leq * < 3$	64.7	96.0	96.0	92.7	92.0	91.3	90.7	94.0
Naturalness								
$1 \leq * < 2$	62.7	27.3	13.3	28.7	24.7	20.0	20.0	6.7
$2 \leq * < 3$	21.3	25.3	20.0	22.7	25.3	19.3	25.3	13.3
$3 \leq * < 4$	12.0	21.3	27.3	27.3	27.3	32.7	31.3	26.7
$4 \leq * \leq 5$	4.0	26.0	39.3	21.3	22.7	28.0	23.3	53.3

Table 7: The percentage of synthetic sentences per augmentation technique falling under each mean opinion score (MOS) range for understandability and naturalness, as obtained through human evaluation.

	LEX_{Dict}	LEX_{Rand}	LEX_{Pred}	EC_{rand}	EC_{spf}	ML_{rand}	ML_{spf}	BT
Understandability	2.16	2.78	2.77	2.72	2.68	2.73	2.70	2.75
Naturalness	1.80	2.84	3.34	2.74	2.84	3.08	2.96	3.76

Table 8: The average mean opinion scores (MOS) for understandability and naturalness per technique, as obtained through human evaluation.