A Morphologically-Aware Dictionary-based Data Augmentation Technique for Machine Translation of Under-Represented Languages

Anonymous ACL submission

Abstract

The availability of parallel texts is crucial to the performance of machine translation models. However, most of the world's languages face the predominant challenge of data scarcity. In this paper, we propose strategies to synthesize parallel data relying on morpho-syntactic information and using bilingual lexicons along with a small amount of *seed* parallel data. Our methodology adheres to a *realistic* scenario backed by the small parallel seed data. It is linguistically informed, as it aims to create augmented data that is more likely to be grammatically correct. We analyze how our synthetic data can be combined with raw parallel data and demonstrate a consistent improvement in performance in our experiments on 14 languages (28 English↔X pairs) ranging from well- to very low-resource ones. Our method leads to improvements even when using only five seed sentences and a bilingual lexicon.[1]

1 Introduction

One of the major challenges in machine translation (MT) is the lack of parallel data for most of the world's languages. Traditional approaches (Wu et al., 2008; Mikolov et al., 2013) used to rely on dictionaries and linguistic knowledge for MT. One of the naive ways to use dictionaries for MT is to translate by looking up words of a source sentence in a bilingual lexicon and replacing their corresponding translations in the target language. However, this approach has certain shortcomings (Wang et al., 2022a). Firstly, the coverage of translations depends on the size and comprehensiveness of the lexicon, which can result in incomplete translations and code-mixed versions of the source and target languages. The translated sentences may also not adhere to the target language's grammatical rules or word order. Furthermore, most dictionaries operate at the lemma level, posing challenges for morphologically-rich languages. Therefore, *solely* relying on dictionaries is not a viable solution for low-resource languages.

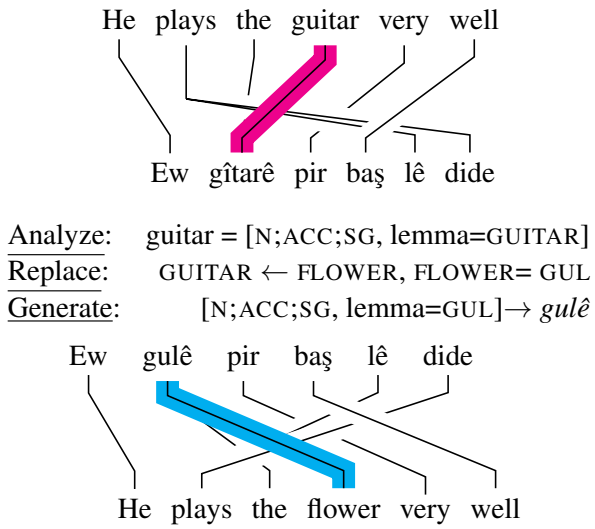

Figure 1: A schema of our approach. After aligning 'guitar' (in English) and '*gîtarê*' (in Kurmanji Kurdish), the new word 'flower' is randomly selected to replace 'guitar' and its translation '*gul*' in a bilingual dictionary is inflected according to its morphological features as '*gulê*'. Small caps refer to lemmata.

In recent approaches to MT that mainly rely on encoder-decoder networks like transformers (Vaswani et al., 2017), the ideal scenario is to train an MT model on a large parallel corpus. Creating a parallel corpus for a given language, however, requires linguistic and technical expertise lacking for under-resourced languages and is also a costly and time-consuming task. To remedy this, recent studies in natural language processing (NLP) focus on unsupervised methods based on monolingual data (Sennrich et al., 2016a; Lample et al., 2018a), back-translation (Edunov et al., 2018a,b), other data augmentation techniques (Sánchez-Cartagena et al., 2021), or fine-tune pre-trained models to adapt to a different language, domain, or dialect (Bapna and Firat, 2019). Therefore, the usage of

[1] Data and code will be publicly released upon acceptance.

dictionaries is largely under-studied, even though they are still practically in use (Peng et al., 2020; Sennrich et al., 2016b).

In this paper, we put forward a dictionary-based approach akin to early dictionary-based MT systems (Tyers, 2009; Koehn and Knight, 2002, 2001; Sánchez-Cartagena et al., 2011) yet more sophisticated as it relies on the morpho-syntactic analysis of words to generate a parallel corpus synthetically. As illustrated in Figure 1, our approach consists of four components: alignment, analysis, replacement, and generation. Given a small set of parallel text as seed data, we first retrieve possible word-level translation pairs in the source and target languages as in 'guitar' and '*gîtar*' in English and Kurmanji Kurdish, respectively. We then morphologically analyze the source words in the translation pairs, e.g., 'guitar' is a singular noun in the accusative case in the example. With the morphological features of a word in the source sentence, we can now sample a word from a bilingual dictionary with the same morphological features, e.g., *gulê*, and "plug" it into our sentence to generate a new sentence pair synthetically. As such, the synthetically-generated sentences are likely to create new grammatically-sound translations.

To summarize, the contributions of our work are three-fold:

- We propose a morphologically-informed replacement method to create a new synthetic sentence.
- We show that this synthetic parallel data helps improve the MT system's quality when mixed with real parallel data.
- We also demonstrate the effectiveness of our method in extremely data-scarce scenarios, where as little as five parallel seed sentences are rendered useful with our approach.

Note that we will interchange the terms "dictionary" and "bilingual lexicon" throughout the paper for readability reasons.

2 Method

Our method requires a small parallel dataset called seed data containing sentences in the source and target languages to create synthetic parallel data. Our approach consists of three components. We first prepare data by tokenizing sentences and obtaining word-level alignment between the parallel sentences. This step is completed by morphologically analyzing aligned word pairs. Then, we replace words considering the morphological features in the augmentation component and filter the synthetic sentences using language models. Finally, we build MT systems in different settings varying the number of synthetic sentences.

2.1 Analysis

Alignment We perform word alignment to our seed data, identifying the relationship between words in the seed sentence. This is necessary for knowing which words are translations of each other. If we replace a word in the source sentence, the aligned target word of the target sentence must also be replaced to reflect the changes.

Morphological Tagging We analyze the entries in the bilingual lexicon's source side words to facilitate the data augmentation process. This way, we can categorize entries based on morphological features and find the part-of-speech (POS) tags, e.g., ADJ, of our bilingual lexicon's source side words.

Word-pair Selection We randomly choose words from the source side, here in English, for each seed sentence. We refer to Figure 1 as our example where we generate the morphological feature and POS tag for the given word "guitar". We also find the translation of "guitar" in the seed sentence's target side. Here, that word is "gîtarê", which we get from the alignment in §2.1. We find the morphological feature and POS tag of "gîtarê" too.

2.2 Augmentation

We introduce two different approaches for the augmentation of the seed sentences:

Morphologically-Informed

1. Referring to Figure 1, we first replace "guitar" with another random word, e.g. "flower", having identical morphological features created in §2.1. As such, a new sentence is synthetically created as "*He plays the flower very well*". It should be noted that this procedure does not consider the semantic relevance of the candidate word. In other words, it may yield nonsensical sentences yet morpho-syntactically valid.

2. Then, we replace "*gîtarê*" with the translation of "flower" being "*gul*" in Kurmanji Kurdish from the bilingual lexicon. It is worth

mentioning that we use PanLex dictionaries[2] where some of the entries are not in the lemma form. Therefore, we also lemmatize the retrieved word form in the dictionary, i.e., "*gul*" to mitigate the impact of the inaccuracy of the lexicographic data.

3. Last, we perform morphological inflection where a lemma is inflected based on morphological features of the word that will be substituted, i.e., '*gîtar*'. Doing this, we create a new sentence where the randomly selected word in the dictionary '*gul*' appears grammatically and morphologically correct as '*gulê*'. We do this to guarantee that the new word follows the correct morphological features. Thus creating a synthetic target translation "*Ew gulê pir baş lê dide*" of the synthetic source sentence "He plays the flower very well".

Naive In contrast to the previous augmentation technique where we consider the morphological features, we carry out a naive random word replacement approach where only the POS tag is identical, without lemmatizing or inflecting the word based on the sentence. For instance in Figure 1, a synthetic sentence created this way would be "He plays the flower very well" and its generated translation "*Ew gul pir baş lê dide*". Here "*gul*" has not been converted into '*gulê*'. In the *Morphologically-Informed* setup, we preserve the morphological information of the word we change, thus making the synthetic data more likely to be grammatically correct. In this *naive* approach, on the other hand, we lose this information.

2.3 Filtering with LMs

We create synthetic sentences for each seed sentence with the above augmentation approaches. Given that the synthetic sentences may not be meaningful, e.g., "He plays the flower very well", we also incorporate information from a language model (LM) by estimating the perplexity (ppl) of the synthetic sentences:

$$ppl(x) = exp\{-1/t \sum_{i}^{t} logp_{\theta}(x_i|x_{<i})\},$$

Where the probability of a sentence of length t containing words x existing in the LM. The lower the perplexity is, the more natural the sentence is.

[2] https://panlex.org/snapshot

We filter the augmented sentences using the LM and rank them based on the perplexity scores to pick the sentences with the correct context. This step yields sentences more likely to appear with the lowest perplexity.

2.4 Neural Machine Translation

Using the synthetic data, we build neural MT systems for each language pair in one direction. To do so, one of the approaches is to train a transformer-based encoder-decoder model from scratch with random weights only on the parallel data. This model type excels in high-resource settings but hardly reaches up to the mark performance for low-resource languages (Duh et al., 2020). Another approach is to fine-tune a model based on a pre-trained model. Instead of initializing with random weights, the training is carried out on a previously-trained transformer model. The pre-trained model can be either monolingual or multilingual and can be pre-trained on any task, normally on denoising ones. This approach (Alabi et al., 2022) is promising to improve low-resource languages as the model does not need to learn all language components from scratch. If the pre-trained model is multilingual, the model can leverage resources from other high-resource languages.

3 Experimental Setup

3.1 Dataset

Parallel Data To create synthetic data, we use the parallel sentences in the OPUS-100 (Zhang et al., 2020) corpus[3] with English as the source language and other languages as target languages. We use this training set as our parallel seed data for training. For testing and validation, we use the `devtest` and `dev` sets of the FLORES-200[4] benchmark (Team et al., 2022) respectively. Table 1 summarizes the statistics of our datasets. We divide the languages into four categories according to their data availability: extremely low-resource, low-resource, well-resourced, and high-resource.

Bilingual Dictionaries We extract dictionaries between English and each target language from the PanLex database containing 25 million words in 2,500 dictionaries of 5,700 languages.

[3] https://data.statmt.org/opus-100-corpus

[4] https://github.com/facebookresearch/flores/tree/main/flores200

Language (code)	# Seed	# Entries
Armenian (HYE)	7,059	161,798
Wolof (WOL)	7,918	4,971
Kurmanji (KMR)	8,199	47,461
Scottish Gaelic (GLA)	16,316	51,416
Marathi (MAR)	27,007	65,559
Uyghur (UIG)	72,170	9,054
Kazakh (KAZ)	79,927	40,516
Tamil (TAM)	227,014	230,882
Irish (GLE)	289,524	71,436
Galician (GLG)	515,344	185,946
Hindi (HIN)	534,319	409,076
Urdu (URD)	753,913	86,695
Greek (ELL)	1,000,000	407,311
Maltese (MLT)	1,000,000	33,131

Table 1: Statistics of our datasets (seed parallel data and dictionary entries). Sorted according to the number of available training sentences.

3.2 Tools

We use `Stanza`[5] for tokenization, morphological feature tagging, POS tagging, and lemmatization. Stanza uses different models for different languages. For word alignment we use `fast_align`[6] (Dyer et al., 2013), and we use `pyinflect`[7] for morphological inflection. Note that `pyinflect` only supports English, but in this work, we only do inflection on the English side for our synthetic data creation framework. We use the HuggingFace (Wolf et al., 2020) toolkit for training the language models.

3.3 Implementation Details

Language Model To construct the language model (LM), we adopt the methodology outlined in the GPT-2 recipe provided by HuggingFace (Radford et al., 2019). We utilize the monolingual side of the parallel data specific to each language as the training dataset. Given that many of the languages involved in our experiment are not highly resourced, we make certain modifications to the GPT-2 model. We employ only six layers instead of the original 12 to mitigate resource limitations. Additionally, we decrease the vocabulary size to 5000. These adjustments help tailor the model to our experiment's specific requirements of low-resource languages.

Seed Data We do not use all the available seed data for creating synthetic sentences. Short sentences with less than seven tokens are not used as seed sentences. As Stanza uses context to generate morphological features, short sentences seem not to provide enough context for the model to produce reasonable annotations.[8]

Lexicons The bilingual lexicons often provide several translations for one source word. We organize the lexicon so that only one translation is available for each source word. We do so naively, only taking the first translation of a word and discarding the rest. We also ensure that the source and target have the same POS tag. To produce the morphological features of the dictionary entries, we rely not only on Stanza[9] but we also perform lookups on *Unimorph*[10] (Batsuren et al., 2022), which provides morphological inflection paradigms for dozens of languages (including the ones we work on) annotated with POS tags and morphological features. For this work, we only work with augmentation, focusing on nouns, adjectives, and verbs.

Synthetic Data We create five sizes of synthetic data: 5K, 10K, 50K, 100K, and 200K for each language pair. In each set, the previous set's data is used. That means that when compiling the 10K synthetic dataset, we create new 5K data to add to the previous 5K data, and so on. This ensures that our experiments only vary based on the newly added synthetic data (and not due to additional randomness).

For each sentence, we randomly choose at most two words for replacement. As the word replacement is random, getting the exact number of sentences for each set is not guaranteed, and there may be duplicates. From each seed sentence, M number of synthetic sentences are created. Let's say we want to make a total of 5,000 seed sentences. Then, M is chosen to get barely more than 5,000 sentences. After that, we sort with the perplexity of the LM and select the sentences with lower perplexity to create that set.

[5] https://stanfordnlp.github.io/stanza/
[6] https://github.com/clab/fast_align
[7] https://pypi.org/project/pyinflect/0.2.0/

[8] This was based on preliminary experiments and manual inspection of Stanza's outputs.
[9] As there is no sentential context, Stanza is bound to be error-prone.
[10] https://unimorph.github.io/

Model Details We fine-tune DeltaLM (Ma et al., 2021), a large pre-trained multilingual encoder-decoder model that regards the decoder as the task layer of off-the-shelf pre-trained encoders. This is done separately for each language, not multilingually. The baseline system is the one that uses only the available real parallel data. Throughout the paper, we refer to the baseline as the *0K (untagged)* model, as it has seen 0 synthetic data. The rest of the models are use tags <*clean*> and <*noisy*> at the beginning of the sentences to distinguish between real and synthetic data. The model's name (e.g. *5K*) indicates how much synthetic data has been added to the seed data during training.

4 Results

From English In 11 out of the 14 Eng-X language pairs, our approaches yield improvements randgng from 0.4 points to more than 3 BLEU (Papineni et al., 2002) points.

Due to space constraint, we show all results for the six language pairs with the largest improvement over the baseline in Figure 2. Comparing the data augmentation methods, our morphologically aware approach yields a better score than the naive one in all cases except for Galician. We find that the augmentation is consistently beneficial for Irish and Galician, regardless of how much data we add. But for other pairs adding more synthetic data does not lead to sustained improvements.

Table A.1 in the Appendix shows our experimental results on all 14 language pairs from English. We use two pre-trained models: *DeltaLM-Base* and *DeltaLM-Large*. We tried to use DeltaLM-Large for all language pairs, but low-resource language pairs quickly overfit on the large model and do not generalize well. Apart from Armenian (HYE), we get a higher BLEU score with our settings for all other language pairs. The improvement margin is negligent in languages where the baseline system is already very bad. Languages like Wolof (WOL) and Uyghur (UIG) have baseline BLEU scores of less than 2, showing us that our parallel seed data is not of good quality. For languages, Kazakh (KAZ), Marathi (MAR), and Tamil (TAM), all rather low-resource languages, the improvement is less than 0.5 BLEU points, but it ranges from 0.79 to 3.21 BLEU score for all other languages. We also observe a similar trend of improvement in the case of adding more noisy data. The score improves to the highest point, but as more synthetic data is added the system gets worse.

To English As before we show the best-performing six language pairs in the X-English direction in Figure 3. Unlike English-X, the patterns here are the same for all language pairs. In all 14 languages except for Armenian our approach improves upon the baseline, and the morphologically-informed method is better than the naive approach. In every case adding more synthetic data after a while does not lead to more improvements.

Table A.2 in the Appendix lists our results on all 14 language pairs in the to-English direction. The BLEU scores are generally higher in this setting, as the pre-trained model has seen a lot of English data on the target side. Apart from Armenian (HYE), we get a higher BLEU score with our settings for all other language pairs, the same as before. The improvement margin is negligent for Greek (ELL) and Maltese (MLT), showing that the language has no room for improvement through this type of augmentation, as the models are already fairly good. For Wolof (WOL) and Uyghur (UIG), our improvement is less than 1 BLEU score. This is also the same as before, showing that the parallel data for these languages is not of high quality. The improvement ranges from 1.12 to 4.24 BLEU scores for all other languages. We also see a similar trend as in the From-English direction; after some point, the more synthetic data we add, the system worsens; most improvements are obtained with 5K and 10K synthetic examples.

5 Analysis

Is the performance tied to any single component? We perform an ablation study to find out which component of the model is responsible for the performance boost. We work on this experiment with five thousand synthetic data and the SCOTTISH GAELIC-ENGLISH direction. We compare three different components of our pipeline:

- **Does the number of the generated synthetic data matter?** For this setup, instead of creating three synthetic sentences from each seed sentence, we create 30 synthetic sentences from each seed sentence. We refer to this setup as a "5K Number".
- **Does the length of the seed sentences matter?** We create synthetic sentences from seed sentences with less than seven tokens for this setup. We refer to this setup as a "5K Length".

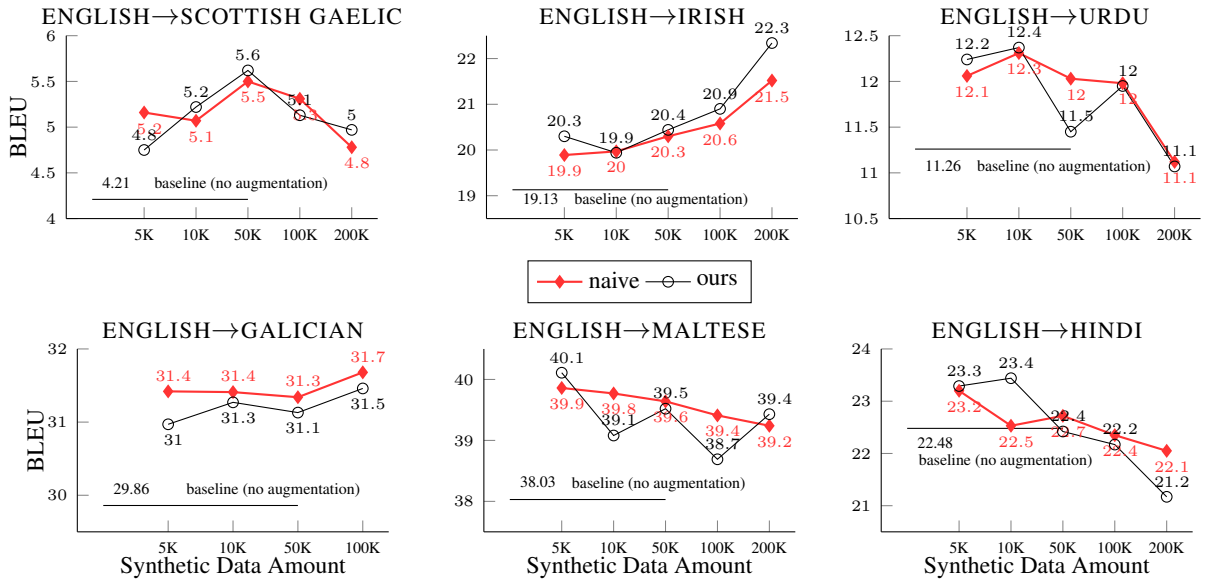

Figure 2: BLEU scores on the test sets for six languages in the ENGLISH-X direction. X-axis indicates the amount of synthetic parallel data we use along with seed data. The baseline uses no synthetic data. Except for Irish and Galician, all the other languages do not benefit from the increasing amounts of synthetic data. It seems like Irish has even room for more improvement. **ours** is the morphologically-informed method.

Ablations on Scottish Gaelic-English		
	Ours	**Naive**
5K	**13.32**	13.05
5K Number	12.76	**13.13**
5K Length	12.58	12.28
5K Align	12.14	12.62

Table 2: Ablation result of the importance of different components of our method. If we don't use one of the components, the BLEU score drops significantly.

- **Does the choice of alignment model play any role?** Instead of aligning the seed sentences with `fast_align`, we use `awesome-align` using "bert-base-multilingual-cased" for this setup. We refer to this setup as a "5K Align".

Table 2 shows the results of these experiments. The main takeaway is that every component of our method is necessary to boost scores: scores decrease when we replace one component. The most important for low-resource languages is to use a compatible alignment model. As large multilingual pre-trained models do not represent them very well, and `awesome_align` relies on such a model, we are better off using `fast_align`. The number of generated synthetic data also matters as we anticipated. The reason is that when we create a huge sentence pool and sample a small number of sentences from there, the number of unique seed sentences that contribute to the synthetic data is reduced. We also confirm that the length of the sentence matters for Stanza: the shorter the sentence is, the less context it has, thus reducing the quality of the morphological analysis and consequently of our synthetic sentences.

Does the number of seed sentences or the amount of new vocabulary matter? To do this experiment, we work again on the GLA-ENG direction to create five thousand synthetic data. We build four different models:

- **5K**: This is the original five thousand-size synthetic dataset we created. We create three sentences from each seed sentence and randomly choose words from all candidate words for replacement.
- **5K (one)**: In this setup, we try to create 5000 sentences from only one seed sentence and randomly choose words from all candidate words for replacement. However, our process could not generate 5,000 unique synthetic sentences from one seed sentence; instead, it took five seed sentences to generate 5,000 synthetic sentences.
- **5K (half)**: In this setup, we create three sentences from each seed sentence and randomly choose words from the first half of the candidate words for replacement.

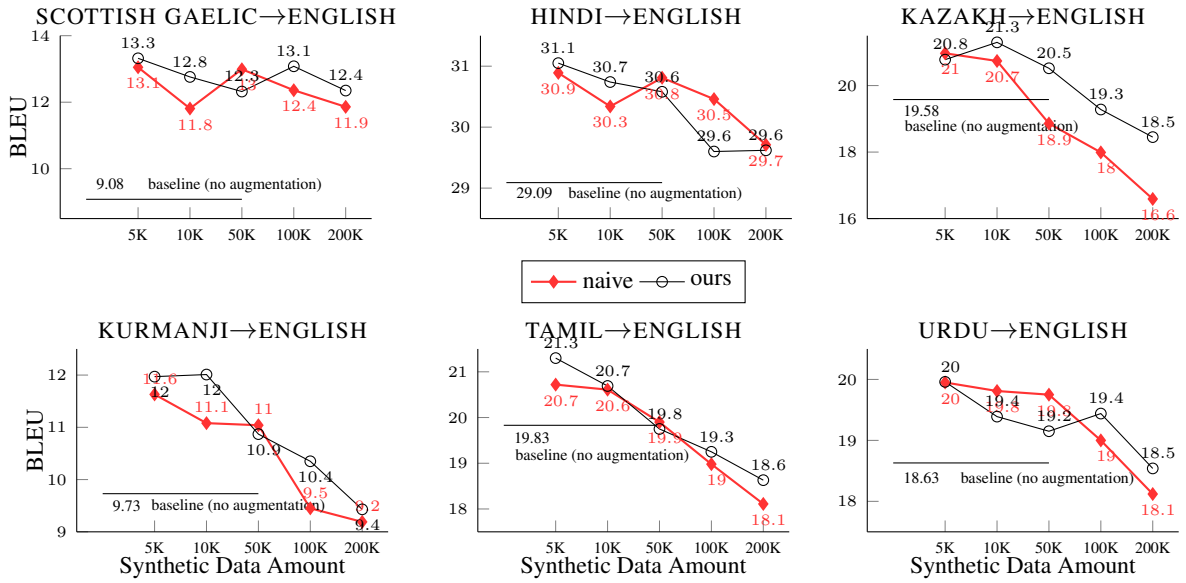

Figure 3: BLEU scores on the test sets for six languages in the X-ENGLISH direction. **ours** is the morphologically-informed method. Morphologically-informed approach outperforms the naive approach in all the six language pairs. X-axis indicates the amount of synthetic parallel data we use along with seed data. The baseline uses no synthetic data.

- **5K (remove)**: In this setup, we create ten sentences from each seed sentence and randomly choose words from all candidate words for replacement, but when we choose a word, we remove that word as a candidate so that it is not chosen again. Ideally, we would have sentences of the same amount of the lexicon vocabulary.

Table 3 shows these experiments' results. An exciting result is the one for 5K (one), where we use only five seed sentences to create five thousand synthetic sentences. In doing so, we introduce 200 new words, but we get a substantial jump of 3.71 BLEU score, which shows the promise of our method. Even if we have few high-quality parallel sentences and a good-quality lexicon, our method is bound to boost MT system quality.

How much does filtering with LM help? To do this experiment, we perform a control experiment to contrast with our filtering-with-LM-perplexity approach. In the control setting, we choose sentences randomly from the pool of synthetic sentences. We randomly select a subset of a hundred thousand seed sentences from the OPUS-100 ENG, ELL dataset and do ablation on both ENG-ELL and ELL-ENG directions.

Table 4 shows these experiments' results. In the random setting, the results are rather unstable, with very low BLEU scores for some settings. This could be because we might be randomly choosing bad sentences from the pool. The results with informed sentence selection (through perplexity), instead, are stable and consistently improving.

6 Related Work

Dictionaries have been and are indispensable resources in various applications in NLP (Wilson et al., 2020; Wang et al., 2019; Xiao and Guo, 2014). More specifically, many previous studies use dictionaries in MT to improve translation quality for low-resource languages with or without monolingual or parallel corpora. A closely related task is bilingual lexicon induction that departs from an unsupervised MT task where no parallel resources, including the ground-truth bilingual lexicon, are incorporated (Artetxe et al., 2017; Lample et al., 2018b). The bilingual lexicon is often utilized as a seed in bilingual lexicon induction that aims to induce more word pairs within the language pair (Mikolov et al., 2013). Another utilization of the bilingual lexicon is for translating low-frequency words in supervised neural MT (Arthur et al., 2016; Zhang and Zong, 2016).

On the usage of dictionaries in MT, Peng et al. (2020) employ dictionaries for cross-lingual MT, Fadaee et al. (2017) propose a data augmentation approach to target low-frequency words by generating sentence pairs containing rare words, Duan et al. (2020) use dictionaries to drive the semantic

	Scottish Gaelic-English Ours				Scottish Gaelic-English Naive			
	BLEU	# ENG types	# GLA types	# seed sentences	BLEU	# ENG types	# GLA types	# seed sentences
0K	9.08	11077	13826	0	9.08	11077	13836	0
5K (one)	12.79	11269	14020	5	12.58	12588	15883	5
5K	**13.32**	12763	15847	1511	13.05	12057	15082	1512
5K (half)	13.13	12367	15216	1527	12.57	11811	14740	1508
5K (remove)	12.91	12528	15702	1909	**13.29**	12511	15694	1909

Table 3: Ablation about the number of new vocabulary introduced and the number of seed sentences used to create five thousand synthetic data. Using as little as five seed sentences boosts the 3.71 BLEU score.

Pairs	0K	5K	10K	50K	100K	200K
English-Modern Greek (random)	14.24	13.98	5.01	**15.7**	14.92	4.23
English-Modern Greek (filtered)		14.62	15.14	15.2	14.99	**15.54**
Modern Greek-English (random)	11.6	12.2	10.24	**18.34**	9.21	12.42
Modern Greek-English (filtered)		16.91	17.25	17.05	**19.01**	17.64

Table 4: Filtering the synthetic data leads to consistent improvements, but random data sampling leads to unstable results. Instead, the BLEU score drops randomly for a random approach.

spaces of the source and target languages becoming closer in MT training without parallel sentences and Wang et al. (2022b) explore the utilization of dictionaries for synthesizing textual or labeled data, focusing on tasks such as named entity recognition and part-of-speech tagging.

Unlike many of the previous approaches that are fixated on only monolingual data, our approach considers using a bilingual lexicon and maintaining morphology in augmentation. Our approach is similar in spirit to Fadaee et al. (2017) technique with additional consideration of morphological complexity in the synthetic data augmentation process. Also, inspired by Wang et al. (2022b)'s approach, our research shares a common thread by using different strategies for synthesizing data using lexicons and integrating such data with monolingual or parallel text when accessible. Both studies aim to leverage lexicons to enhance various NLP tasks, albeit in different contexts.

7 Conclusion

Our approaches have proven beneficial for most of the 14 languages under investigation, except for Armenian. Even if the improvements in BLEU scores may be small for some languages, there is a noticeable boost in most. Interestingly, we observed improvements even in language pairs (e.g., WOL-ENG, ENG-KMR, ENG-GLA) with unsatisfactory initial baseline scores. This observation suggests our approach can enhance performance even in more challenging scenarios. The results also highlight the importance of obtaining high-quality seed sentences. We found that as few as five good-quality seed data points can contribute to creating five thousand synthetic data samples of good quality that would boost performance. This data augmentation process could play a vital role in improving the overall performance of machine translation systems and be combined with other augmentation techniques (e.g., back-translation) as we deem it orthogonal to them.[11]

Future Work In our current work, we focused on conducting morphological inflection exclusively on the English side of the translation task. The main reason for this choice was the availability of a reliable morphological inflector specifically designed for English. However, we encountered challenges when applying the same approach to other languages. We lacked suitable morphological inflection tools for those languages, or the accuracy of the available tools did not meet our requirements. Incorporating these tools would have posed a significant bottleneck to the effectiveness of our approach. For future research, we aim to explore how our approach can be extended by performing morphological inflection in other languages. This involves developing or obtaining accurate and reliable morphological inflection tools.

[11] We note that back-translation is rarely effective in most extremely low-resource languages due to the abysmal quality of the initial systems.

8 Limitations

One of the limitations of our current approach is the use of the Stanza model. Since bilingual lexicons have no context, relying solely on morphological features obtained from the lexicon results in more general features. This can be particularly challenging in morphologically rich languages, where a single word can have multiple meanings depending on the sentence context. Another limitation is the language support provided by the Stanza model, which is currently limited to 60 languages. This constraint restricts the applicability of our approach to only those languages supported by Stanza. To expand our work to languages not supported by Stanza, it is necessary to create custom Stanza models specifically tailored for those languages. This process requires additional time and effort to develop and validate the models for each language of interest.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1538–1548. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1570–1579. Association for Computational Linguistics.

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. 2020. Benchmarking neural and statistical machine translation on low-resource African languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2667–2675, Marseille, France. European Language Resources Association.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018a. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*,

pages 489–500. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018b. Understanding back-translation at scale.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2001. Knowledge sources for word-level translation models. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

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

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. 2020. Dictionary-based data augmentation for cross-domain neural machine translation. *CoRR*, abs/2004.02577.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8502–8516. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2011. Integrating shallow-transfer rules into phrase-based statistical machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Francis M. Tyers. 2009. Rule-based augmentation of training data in Breton-French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Qi Wang, Yangming Zhou, Tong Ruan, Daqi Gao, Yuhang Xia, and Ping He. 2019. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *Journal of biomedical informatics*, 92:103133.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022a. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022b. Expanding pretrained models to thousands more languages via lexicon-based adaptation.

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4764–4773.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Hua Wu, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 993–1000, Manchester, UK. Coling 2008 Organizing Committee.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 119–129. ACL.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Bridging neural machine translation and bilingual dictionaries. *CoRR*, abs/1610.07272.

A Appendix

Pairs	0K (untagged)	5K (tagged)	10K (tagged)	50K (tagged)	100K (tagged)	200K (tagged)	Δ
			DeltaLM-Base				
ENG-GLA Naive	4.21	5.16	5.07	5.5	5.31	4.78	1.41
ENG-GLA Ours		4.75	5.22	**5.62**	5.13	4.97	
ENG-HYE Naive	**5.61**	4.86	4.65	2.91	2.75	2.11	0.0
ENG-HYE Ours		4.28	4.52	3.3	2.97	2.83	
ENG-KAZ Naive	6.13	6.34	5.93	5.16	5.39	4.58	0.37
ENG-KAZ Ours		6.28	**6.5**	5.27	5.68	4.79	
ENG-KMR Naive	1.66	2.36	2.05	2.12	1.66	1.82	0.79
ENG-KMR Ours		**2.45**	2.27	2.32	1.89	1.78	
ENG-WOL Naive	1.08	0.78	1.03	1.19	1.05	0.94	0.12
ENG-WOL Ours		1.08	1.13	1.11	**1.2**	1.02	
			DeltaLM-Large				
ENG-ELL Naive	23.06	22.2	**23.09**	22.17	22.55	23.01	0.03
ENG-ELL Ours		22.17	21.93	22.24	22.64	23.07	
ENG-GLE Naive	19.13	19.89	19.97	20.3	20.58	21.52	3.21
ENG-GLE Ours		20.3	19.94	20.44	20.9	**22.34**	
ENG-GLG Naive	29.86	31.42	31.41	31.34	**31.68**		1.82
ENG-GLG Ours		30.97	31.27	31.13	31.46		
ENG-HIN Naive	22.48	23.2	22.53	22.72	22.35	22.05	0.96
ENG-HIN Ours		23.29	**23.44**	22.42	22.17	21.17	
ENG-MAR Naive	5.03	4.86	4.64	4.55	4.54	4.04	0.44
ENG-MAR Ours		**5.47**	5.35	4.76	4.72	4.34	
ENG-MLT Naive	38.03	39.86	39.77	39.64	39.41	39.24	2.08
ENG-MLT Ours		**40.11**	39.08	39.52	38.69	39.43	
ENG-TAM Naive	5.3	5.64	5.33	5.35	5.10	5.58	0.46
ENG-TAM Ours		5.44	**5.76**	5.28	5.48	5.34	
ENG-URD Naive	11.26	12.06	12.31	12.03	11.98	11.12	1.11
ENG-URD Ours		12.24	**12.37**	11.45	11.95	11.07	
ENG-UIG Naive	1.24	0.7	0.88	0.76	0.87		0.04
ENG-UIG Ours		0.63	1.06	1.17	**1.28**		

Table A.1: BLEU score of 9 languages from ENG-X direction. Columns indicate the amount of synthetic parallel data we use. The **0K (untagged)** column is our baseline. The rows indicating **Naive** is the approach where we replace words of the same POS tag. The rows indicating **Ours** is the approach where we replace words of the same morphological feature. Δ is the difference between the baseline and the best model's score. Δ is 0.0 if the baseline is the best model.

Pairs	0K (untagged)	5K (tagged)	10K (tagged)	50K (tagged)	100K (tagged)	200K (tagged)	Δ
			DeltaLM-Base				
HYE-ENG Naive	**16.04**	15.78	13.84	9.71	8.39	7.79	0.0
HYE-ENG Ours		15.59	14.63	10.11	9.24	8.72	
WOL-ENG Naive	1.83	**2.59**	2.24	2.09	1.53	1.21	0.76
WOL-ENG Ours		2.31	2.22	1.71	1.55	1.1	
			DeltaLM-Large				
GLA-ENG Naive	9.08	13.05	11.81	12.99	12.36	11.86	4.24
GLA-ENG Ours		**13.32**	12.76	12.32	13.08	12.35	
KAZ-ENG Naive	19.58	20.97	20.74	18.86	17.99	16.59	1.72
KAZ-ENG Ours		20.78	**21.3**	20.52	19.28	18.45	
KMR-ENG Naive	9.73	11.63	11.08	11.04	9.45	9.19	2.28
KMR-ENG Ours		11.97	**12.01**	10.87	10.35	9.43	
ELL-ENG Naive	31.94	32.33	32.32	32.34	32.35	31.98	0.42
ELL-ENG Ours		**32.36**	31.88	32.34	32.23	32.1	
GLE-ENG Naive	28.71	30.03	29.54	28.65	28.73	28.57	1.32
GLE-ENG Ours		29.96	29.7	**30.03**	28.99	29.36	
GLG-ENG Naive	37.07	37.89	38.01	37.89	38.02		1.12
GLG-ENG Ours		37.84	**38.19**	38.02	37.61		
HIN-ENG Naive	29.09	30.89	30.34	30.81	30.46	29.71	1.96
HIN-ENG Ours		**31.05**	30.74	30.58	29.6	29.62	
MAR-ENG Naive	22.96	23.66	23.36	22.1	21.22	20	1.15
MAR-ENG Ours		**24.11**	23.54	22.05	21.78	19.92	
MLT-ENG Naive	45.21	45.2	45.43	45.23	45.23	44.9	0.44
MLT-ENG Ours		**45.65**	45.07	44.87	44.94	44.7	
TAM-ENG Naive	19.83	20.72	20.61	19.9	18.98	18.11	1.47
TAM-ENG Ours		**21.3**	20.69	19.75	19.25	18.63	
URD-ENG Naive	18.63	19.95	19.81	19.75	19	18.12	1.33
URD-ENG Ours		**19.96**	19.39	19.15	19.44	18.54	
UIG-ENG Naive	10.49	11.08	11.27	10.67	9.56	8.77	0.83
UIG-ENG Ours		**11.32**	11.31	10.71	9.76	8.18	

Table A.2: BLEU score of 9 languages from X-ENG direction. Columns indicate the amount of synthetic parallel data we use. The **0K (untagged)** column is our baseline. The rows indicating **Naive** is the approach where we replace words of the same POS tag. The rows indicating **Ours** is the approach where we replace words of the same morphological features. Δ is the difference between the baseline and the best model's score. Δ is 0.0 if the baseline is the best model.