Improving Translation of Out Of Vocabulary Words using Bilingual Lexicon Induction in Low-Resource Machine Translation

Jonas Waldendorf jonas.waldendorf@ed.ac.uk
Alexandra Birch a.birch@ed.ac.uk
Barry Haddow bhaddow@inf.ed.ac.uk
Antonio Valerio Miceli Barone amiceli@ed.ac.uk
School of Informatics, University of Edinburgh

Abstract

Dictionary-based data augmentation techniques have been used in the field of domain adaptation to learn words that do not appear in the parallel training data of a machine translation model. These techniques strive to learn correct translations of these words by generating a synthetic corpus from in-domain monolingual data using a dictionary obtained from bilingual lexicon induction. This paper applies these techniques to low resource machine translation, where content distribution is often shifted between the parallel data and any monolingual data. English-Pashto machine translation systems are trained using a novel approach that introduces monolingual data to existing joint learning techniques for learning bilingual word embeddings, combined with word-for-word back-translation to improve the translation of words that do not or rarely appear in the parallel training data. Improvements are made in terms of BLEU, chrF and word translation accuracy for an En→Ps model, compared to a baseline and when combined with back-translation.

1 Introduction

One difficulty of low-resource neural-machine translation (NMT) is the ability of models to correctly predict words that are out of vocabulary (OOV). OOV words are of particular interest when working with low-resource language pairs as such pairs generally exhibit a more significant shift in the distribution of content between the training and test data compared to high-resource language pairs. The available training data for low-resource languages often contains a significant amount of content from specific domains such as IT and religious texts (Tiedemann, 2012) which is not the case in common down-stream tasks for NMT systems such as translating news articles. Hence, the task of improving the prediction of OOV words in low-resource NMT has significant benefits when deploying such models in realistic inference scenarios. Additionally, the overall low amount of parallel data inherent in the task means that the vocabulary covered by the training data is naturally smaller than the vocabulary covered in a more well-resourced NMT scenario. This work aims to improve the prediction of target side OOV words for an English-Pashto (En-Ps) NMT system. To improve the translation of OOV words we incorporate monolingual target-side data when training NMT models by generating synthetic source-side sentences.

Incorporating monolingual target-side data using back-translation (BT) (Sennrich et al., 2016) has been shown to improve the overall performance of low-resource NMT systems in

terms of automatic sentence-level evaluation metrics such as BLEU or chrF. However, an important benefit of incorporating monolingual data is the increase in the amount of vocabulary that is observed during training, the effects of which can only be seen when evaluating NMT predictions at the level of single OOV words. Work in the field of domain adaptation has shown that word-for-word (WFW) back-translation using bilingual dictionaries extracted from bilingual word embeddings (BWE) is a suitable alternative to BT when specifically targeting improved translation of OOV words (Hu et al., 2019). Whilst the source-side sentences produced by WFW-BT have lower adequacy and fluency, the key benefit compared to BT is that they more frequently result in direct supervision of OOV words. That is to say, WFW-BT more frequently results in source-side sentences that contain a correct translation of target side OOV words and, by extension, improve the ability of the NMT model to predict those words correctly. Inspired Hu et al. (2019) we adopt the WFW-BT methodology to improve the prediction accuracy of OOV target side words for the En-Ps NMT model.

Compared to the dictionary-based techniques in domain adaptation (Hu et al., 2019; Huck et al., 2019), which aim to predict target-side words specific to the domain correctly, our goal is to correctly predict items from the more varied monolingual vocabulary which are not present in the restricted parallel vocabulary. The consequence of this is that the task is not targeted at a specific set of vocabulary, and by extension, examples of OOV words are less frequent. Additionally, OOV words are less likely to appear in similar contexts on both sides of the monolingual data because we rely on the assumption that the distribution of content is the same across languages. This assumption only holds weakly for English and Pashto, which are both linguistically and culturally distinct (Shen et al., 2021). Moreover, the morphological complexity of Pashto means that many words have considerably more surface forms than their English counterparts, all of which should all translate to the same English word.

As a result of the above observations obtaining a bilingual embedding space (BWE) that correctly maps not only frequent words but specifically OOV words is challenging. We propose a new approach to obtaining BWEs based on the findings of Søgaard et al. (2018); Ormazabal et al. (2019) that joint training (Luong et al., 2015) leads to more isomorphic BWE spaces for linguistically distinct languages. However, joint training requires parallel data to train and hence only maps the embedding spaces of words in the parallel data. Our approach trains on the parallel data using joint training to anchor an embedding space whilst simultaneously training on monolingual data. In addition, we incorporate sub-word information into the joint trai - ‘v bm,ning approach as we hypothesise that sub-word information will help alleviate the data sparsity due to Pashto’s morphological complexity. The main contributions of this work are as follows:

- Adapting the WFW-BT methodology to the genuine low-resource scenario of En-Ps NMT to improve the prediction of OOV target side words. This work contributes to the wider task of expanding the often more limited vocabularies of low-resource NMT systems.
- Proposing an extension to the joint training methodology of Luong et al. (2015) that simultaneously trains on monolingual data, to obtain a stronger BWE space.

2 Related Work

Hu et al. (2019); Huck et al. (2019); Peng et al. (2020) all use dictionary-based methods for data augmentation in a domain adaptation setting focusing on OOV words. Hu et al. (2019); Huck et al. (2019) both use bilingual lexicon induction (BLI), but do so in a high resource setting with artificial monolingual data, which is generated by selecting alternating sentences from a parallel corpus as monolingual data. Peng et al. (2020) make use of a high quality pre-existing dictionary to learn new translations. Our work also uses BLI and WFW-BT; however, we apply

these methods to a genuine low-resource NMT problem. Rather than learning the translations of a specialised subset of vocabulary from monolingual data that contains these words on both sides, we are trying to train NMT models to correctly predict words outside the more specialised vocabularies often found in low-resource parallel data sets.

WFW translations also play an important role in unsupervised NMT (UNMT), where they are used to bootstrap NMT models (Artetxe et al., 2019; Lample et al., 2018) before applying iterative BT. However, UNMT requires careful model choices, works poorly when languages have low amounts of monolingual data (Guzmán et al., 2019) and neglects the fact that there is often a small amount of parallel data available. Additionally, it does not focus on the correct prediction of OOV words but rather on the sentence-level translation quality. This work directly uses the available parallel data to train NMT models and uses the monolingual data to improve the prediction of OOV words.

Mapping-based approaches have been the dominant methodology for BLI, reporting strong results whilst only requiring weak supervision or no supervision by using discriminator networks or identical tokens (Conneau et al., 2017; Artetxe et al., 2018). These methods are based on the assumption of isomorphism between word embedding spaces (Søgaard et al., 2018) and require sufficient monolingual data to learn semantically meaningful word embeddings (Artetxe et al., 2020). Luong et al. (2015) propose joint training of BWE spaces using automatically extracted word alignments as a parallel signal and Ormazabal et al. (2019) observe that joint training leads to increased isomorphism. Eder et al. (2020) introduce an anchor-based method to improve BLI from low-resource language pairs. This work builds on Luong et al. (2015) by incorporating monolingual data and sub-word information to learn stronger BWE spaces. Unlike other BLI tasks, which are often only evaluated on words that appear relatively frequently, we are specifically interested in the BLI performance on less frequent words.

Liu et al. (2020) proposed mBart, a masked language model (MLM) sequence-to-sequence pre-training that aligns the token level representations across many languages. Along with BT (Sennrich et al., 2016) MLM pre-training is the most common way of incorporating monolingual data. We incorporate a mBart-like[1] methodology when training our NMT models to ensure a strong baseline. Vulić et al. (2020) find that for low-resource languages static word embeddings perform better than MLM on BLI tasks which they attribute to a better lexical alignment. Based on this Chronopoulou et al. (2021) combine MLM with BWEs to initialise UNMT models. Whilst the aim of our work is different, these results demonstrate that BWEs are still a suitable tool for learning alignments between lexical items and, by extension, improving the prediction of OOV words. Finally, mRASP (Lin et al., 2020) is an alternative to MLM whereby words and phrases are brought into a similar representation space by substituting aligned words in parallel data sets using dictionaries. mRASP is more closely linked to our work than mBart as it focuses on introducing aligned words during pre-training.

3 Methodology

Our approach is split into two distinct stages. The first is obtaining a pseudo-parallel corpus, and the second is training NMT models using the corpus. Below we outline the approaches used to obtain a BWE space by combining joint training with monolingual data, extracting a dictionary from the BWEs and how the dictionary is used to translate target-side monolingual data. Together these three steps represent the WFW-BT methodology which is used to obtain the pseudo-parallel corpus. When learning the bilingual embedding spaces, sub-word information is incorporated either by using FastText (Bojanowski et al., 2017) for mapping-based approaches or by representing words as a combination of n-grams in the same manner as FastText for Bivec approaches.

[1] https://github.com/Avmb/marian-mBART

3.1 Mapping

In a mapping-based approach, two sets of monolingual embeddings are trained independently before the embeddings are mapped into the same vector space. Conneau et al. (2017) provide both unsupervised and supervised methods for mapping. However, due to the linguistic dissimilarities between English and Pashto, the mapping baseline focuses on the supervised approach (unsupervised training obtained no correct translations). The supervised approach uses a small seed dictionary to induce the mapping, which is extracted from automatic alignments. The embedding spaces are mapped by iteratively solving the Procrustes problem for the seed dictionary before extracting a new dictionary of nearest neighbours.

3.2 Bivec

The joint training methods are all inspired by Bivec, the approach first introduced by Luong et al. (2015). In comparison to the mapping-based approach, Bivec incorporates a bilingual signal into the loss. For languages l_1 and l_2 this can be viewed as training four skip-gram models simultaneously in the following directions $l_1 \rightarrow l_1$, $l_2 \rightarrow l_2$, $l_1 \rightarrow l_2$ and $l_2 \rightarrow l_1$. Models that train on both languages take word alignments as input, so Bivec can only train on parallel data. If a given word w_1 in l_1 is aligned with another word w_2 in l_2 then w_1 is used to predict the context of w_2 and vice versa.

$$loss = \alpha * (Mono_1 + Mono_2) + \beta * Bi_1 \tag{1}$$

During training updates occur for parallel sentence pairs according to Equation 1, where $Mono$ is the monolingual loss for a sentence, Bi is the bilingual loss for a sentence pair, where α and β are hyperparameters. Luong et al. (2015) utilise Word2Vec (Mikolov et al., 2013) in order to jointly train Skipgram models for l_1 and l_2.

3.3 Bivec with Monolingual Data

We hypothesise that we can anchor the embedding space using joint learning over the parallel data while simultaneously training on the monolingual data so that words that only appear in the monolingual data are also contained in the embedding space. We test the following methods for using monolingual data in bivec.

Bivec Para: For the baseline approach, we initialise the embedding tables with both the parallel and monolingual vocabularies whilst only training on the parallel data. As words are represented as a combination of n-grams to incorporate sub-word information, two similar words (for example perfect/imperfect marking of a verb with a suffix in Pashto) should share many of the same n-grams. Hence, there is a degree of transfer learning if one of the forms is present in the parallel corpus. The primary purpose of this baseline is to establish whether subsequent improvements are due to this inherent transfer learning or from incorporating the monolingual data more directly.

Bivec MonoPost: In this approach we train a Bivec Para model initially to anchor the embedding space. Subsequently, we train on just the monolingual data with no parallel signal to try and learn translations of the monolingual data.

Bivec MonoPre: This approach is the inverse of Bivec MonoPost, first training on just the monolingual data and then training with the Bivec approach on the parallel data. The motivation is to first learn good embedding spaces for each language independently before using Bivec to move the embeddings into the same vector space.

Bivec Combined: Combined training incorporates the monolingual data into the parallel training. In the baseline approach, each iteration updates the model in all four directions. The combined approach adds an additional update for the $l_1 \rightarrow l_1$, $l_2 \rightarrow l_2$ directions using only sentences from the monolingual data.

$$loss = \alpha * (Mono_1 + Mono_2) + \beta * Bi_1 + \gamma * JMono_1 + \delta * JMono_2 \quad (2)$$

This is formalised in Equation 2, where $JMono$ is the loss for the monolingual sentences, $Mono$ is the monolingual loss for the parallel sentences and γ and δ are hyper-parameters. The hyperparameters allow the loss to be adjusted to account for varying amounts of data in the two monolingual corpora as well as the parallel corpus.

3.4 BLI and Word-for-Word Translation

To generate a noisy pseudo-parallel corpus from the monolingual data the approach of Hu et al. (2019) is adopted. First, a bilingual lexicon is extracted from BWEs, and then target-side monolingual sentences are translated word-for-word using the dictionary. The lexicon is extracted using the CSLS (Cross Domain Similarity Scaling) distance metric first introduced by Conneau et al. (2017) to find nearest neighbours. Each word w_1 in l_1 and its nearest neighbour in l_2, w_2 are added to the lexicon if w_1 also appears in the top n nearest neighbours of the w_2 word. Lexicons are extracted separately in the $l_1 \rightarrow l_2$ and the $l_2 \rightarrow l_1$ directions. We translate monolingual target-side data word-for-word using the extracted lexicons. If a word does not appear in the lexicon the target-side token is copied into the translation. We refer to such pseudo-parallel copora as WFW-BT.

4 Experimental Design

The experimental setup is chosen to investigate a genuine low-resource language paired with English.

4.1 Training Data

The data used is adopted from Birch et al. (2021)'s data and as such the initial parallel data is the WMT 2020 data excluding Paracrawl (Barrault et al., 2020). Additional parallel data is provided by the En-Ps corpus from the ByteDance team (Koehn et al., 2020; Xu et al., 2020). The monolingual Pashto data was taken from the Pashto NewsCrawl release[2] and the English monolingual data was taken from the 2019 English NewsCrawl release[3], however only the first 5 million sentences of the English data are used as Birch et al. (2021) report no improvements when using more. The monolingual Pashto data also includes additional crawled sentences from (Birch et al., 2021).

Dataset	No. Sentences
WMT - Parallel	123,198
ByteDance - Parallel	440,000
NewsCrawl - Ps	760,379
NewsCrawl - En*	5,000,000
Crawled - Ps	589,864

Table 1: Number of sentences in corpora used to train BWE's and NMT systems. *Only the first 5,000,000 sentences were used from English NewsCrawl release.

All BWEs are trained with all the available data shown in Table 1. The initial NMT models were trained with both the WMT parallel data and the ByteDance corpus. Only mBart pretraining utilises the full English NewsCrawl corpus; any back-translations, either WFW or using NMT systems, only use the first 5 million monolingual English sentences. For both English and Pashto, the corpora are preprocessed using cleaning and punctuation normalisation scripts from the Moses[4] toolkit (Koehn et al., 2007). For the BLI task, English corpora are lower-cased, and

[2] http://data.statmt.org/news-crawl/ps/

[3] http://data.statmt.org/news-crawl/en/news.2019.en.shuffled.deduped.gz

[4] https://github.com/marian-nmt/moses-scripts

all punctuation is removed for both languages. Data for NMT is tokenised using scripts from Moses before sub-word tokenisation is performed using SentencePiece[5] (Kudo and Richardson, 2018), with a vocabulary size of 16,000. Note that all word-level evaluation metrics first perform the BLI preprocessing steps on the NMT output.

4.2 Test Data

The WMT test set as well as the BBC test set (the combined development and test sets from Birch et al. (2021)) are held out for evaluating both the BLI task and the NMT models, whilst the WMT dev set was used for early stopping when training the NMT models. The BBC test set comprises 2350 sentences from BBC news articles.

4.3 OOV and Rare Words

OOV words are defined as words that do not occur in the parallel data but are present in the BBC test set. We limit these words further by ensuring they are present in the monolingual data. Specifically, as all embeddings are learnt for words that appear ≥ 5 times in a given corpus, OOV words are taken to be words that are not in the parallel data and occur ≥ 5 times in the monolingual data. To expand the analysis we also report results on rare words. Rare words are defined as those words that are not common in the parallel data and are grouped by the frequency with which they appear in the parallel data. Table 2 shows the number of OOV and rare words appearing in the BBC test set for English and Pashto and defines the frequency-based bins for rare words. No OOV words are explicitly mined from the WMT test set.

Word Frequency	En	Ps
0 (OOV)	521	727
1-5	642	1021
6-10	389	449
11-15	295	289
16-20	199	275
21-25	205	186

Table 2: Number of OOV and rare words in the BBC test set at for each word frequency bin. The frequencies refer to the number of times a word appears in the parallel data.

4.4 Bilingual Word Embeddings

As Bivec (Luong et al., 2015) is an extension of Word2Vec, or in the sub-word unit case FastText, the standard hyperparameters are kept constant and are in line with previous work (Søgaard et al., 2018; Ormazabal et al., 2019). Embeddings are 300 dimensional and trained using skip-gram with negative sampling. The minimum word count is set to 5 occurrences across both parallel and monolingual corpora. Models are trained with a learning rate of 0.025, a window size of 10, 10 negative samples and a sampling threshold of 10^{-4}.

Mapping based approaches are trained using the MUSE[6] library (Conneau et al., 2017; Lample et al., 2017). FastAlign[7] (Dyer et al., 2013) alignments are obtained using default settings over 10 iterations on the parallel training data. The seed dictionary for MUSE is extracted using these alignments; the 5000 most frequent Pashto words and the aligned English words are used as a seed dictionary. Similarly, the Pashto words in the frequency range 5000-6500 are used as a validation dictionary when training MUSE.

All FastText embeddings for MUSE are trained for 5 epochs, whereas the combined Bivec model is trained for 20 epochs. Note that the combined Bivec model loops over the parallel datasets, whereas for the FastText embeddings the loop is over all sentences. For the combined Bivec model the learning rate hyperparameters in Equation 2 are α is 0.2, β is 2, γ is 0.5, and

[5] https://github.com/google/sentencepiece
[6] https://github.com/facebookresearch/MUSE
[7] https://github.com/clab/fast_align

δ is 0.2, where language one is English. These values were selected empirically based on the collected evaluation dictionary introduced below. However, this is not an exhaustive sweep of parameters. When extracting the bilingual lexicon using the CSLS metric, the nearest neighbour parameter is set to 5 for En→Ps and 10 for Ps→En. The value of n was selected empirically to provide similar coverage of the vocabulary in both directions.

4.5 Neural Machine translation

Using the Marian Toolkit (Junczys-Dowmunt et al., 2018), the NMT models were trained using the transformer-base alias. Early stopping was performed after ten epochs on the WMT validation set using the mean cross-entropy loss. Models are first trained using an mBART (Liu et al., 2020) like objective on the entire data, which pre-trains the model using the same denoising objective as mBart but only on English and Pashto data. We trained systems using only the parallel data and using pseudo-parallel corpora from BT as a comparison to the WFW-BT based methods. All pseudo-parallel corpora up-sample the parallel data so that there is an approximately equal split of genuine and pseudo-parallel data. Below is a summary of the systems trained:
Baseline: The baseline system is trained only on the parallel WMT and ByteDance data.
WFW-Bivec: Uses a pseudo-parallel corpus generated from WFW-BT using a dictionary obtained with the Bivec Combined methodology and the parallel data.
WFW-MUSE: Uses a pseudo-parallel corpus generated from WFW-BT translation using a dictionary obtained with the MUSE methodology and the parallel data.
BT: Uses a pseudo-parallel corpus generated with back-translation from the Baseline model in the opposite translation direction.
BT-from-WFW: Uses a pseudo-parallel corpus generated with back-translations from the WFW-Bivec model in the opposite translation direction.

4.6 Evaluation Metrics

As there are no freely available, machine readable dictionaries for Ps-En, a small dictionary of 1000 words was collected, which is referred to as the Parallel Dictionary. This dictionary is informed by the FastAlign alignments from the parallel training data that are outside the 6,500 most common Pashto words and are verified using online resources. A second smaller dictionary of 200 words is extracted from the BBC test set by manually aligning Pashto OOV words to their English translations using online translation tools. This dictionary is referred to as the BBC Dictionary. Both lexicons are used to evaluate the different methods of obtaining BWEs using a BLI task translating from Pashto to English.

Sentence-Level Accuracy: Complementing BLI metrics, the BBC test set is directly used to assess the performance of the extracted dictionaries for OOV and rare words. For a given target-side word, all sentences in which it appears are collected from the BBC test set. Then a positive example is defined if at least one of the corresponding source-side sentences contains the correct translation of the target-side word according to the dictionary. Accuracy is reported in both translation directions, based on whether or not a translation was found. In addition, as it is an automatic metric, it is also used to evaluate performance for rare words at each frequency.

Evaluating NMT: The NMT systems are evaluated using BLEU and chrF calculated using SacreBLEU (Post, 2018). In addition, to evaluate the performance on OOV and rare words at the word level we report the micro-averaged F1 score. Each reference translation that contains an OOV or rare word is compared to a given prediction to see if it contains the same token to calculate the F1 score.

5 Results

Table 3 gives the results for the Ps→En BLI task using the dictionaries described in Section 4.6. As expected, the introduction of joint training in the Bivec-based models improves the precision on the Parallel Dictionary, which is comprised of words predominantly present in the parallel data, compared to the precision of the MUSE baseline. Although the Bivec Combined methodology has the best precision, it only slightly improves upon the MUSE baseline on the BBC dictionary. In combination with the overall low results on the BBC dictionary, this highlights the task's difficulty. The low performance is attributed to the comparatively low amount of Pashto data and low frequency of OOV words in the monolingual data. The median number of counts for Pashto OOV words in the monolingual data is 19, which means that the distribution of the contexts in the sample is unlikely to represent the true distributions of contexts in the entire population. In addition, BLI is based on the underlying assumption that the used corpora are at least comparable; that is to say, their distributions are at least similar. This likely holds to some extent for the parallel data, but there are likely significant differences between the monolingual corpora.

Name	Precision Parallel Dictionary			Precision BBC Dictionary		
	@1	@5	@10	@1	@5	@10
Bivec MonoPost	17.56	33.56	40.44	6.63	21.43	27.04
MUSE	24.38	38.42	43.16	9.62	24.06	30.48
Bivec Para	40.62	53.53	60.59	11.22	21.94	24.49
Bivec MonoPre	40.62	55.42	60.76	8.60	15.05	18.82
Bivec Combined	**44.39**	**57.54**	**64.21**	**12.83**	**26.73**	**32.62**

Table 3: Table of the precision at 1, 5 and 10 (top nearest neighbours) for the BLI task in the Ps→En direction for the Parallel and BBC dictionaries.

Contrasting the Bivec Combined methodology to the Bivec Para baseline reveals that just training on parallel data with joint training and solely relying on sub-word information to translate unseen words achieves similar performance to incorporating monolingual data directly especially on the precision @1 metric for the BBC dictionary. This result demonstrates that the sub-word information is critical for learning OOV translations. However, for the Parallel Dictionary, Bivec Combined achieves higher precision values than Bivec Para, suggesting that while the translation of OOV words remains challenging, the introduction of the monolingual data does improve the overall quality of the BWE space.

Word	Combined		MUSE	
Frequencies	Ps	En	Ps	En
0	5.70	2.26	8.36	4.52
1-5	11.05	7.63	10.75	7.33
6-10	14.19	10.75	13.29	10.48
11-15	18.37	10.81	17.01	9.27
16-20	25.19	21.50	19.63	17.29
21-25	23.64	15.52	20.61	9.77

Table 4: Sentence-Level Accuracy metric at each frequency for MUSE and Bivec Combined. The language tag specifies the target language from which sentences are collected.

Compared to the BLI results discussed above, the sentence-level accuracy results given in Table 4 paint a slightly different picture. Although for frequencies of 5 and above Bivec Com-

bined achieves higher accuracy, MUSE obtains a higher accuracy for OOV words. The sentence accuracy metric is noisier than BLI; for example, MUSE translates one of the Pashto OOVs to "him" instead of "regret". As the source-side sentence contains both "him" and "regret" this is still counted as a positive result. However, a qualitative evaluation of the correctly translated Pashto OOV words supports the finding that MUSE correctly translates a higher proportion of the OOV words. Finally, the translation accuracies for English OOV words are lower than those for Pashto at all frequencies, which we attribute to Pashto's higher morphological complexity.

Table 5 gives the BLEU and chrF scores for the En-Ps models for the WMT and BBC test set. Compared to the baseline, WFW-Bivec shows a slight improvement on the BBC test set. Significantly it outperforms WFW-MUSE on both test sets, and in fact, WFW-MUSE appears to decrease performance on the WMT test set compared to both the Baseline and WFW-Bivec. As expected back-translation outperforms both WFW based methods, this seems reasonable as any synthetic source-side sentences generated by back-translation are likely to be more fluent, especially as Pashto and English exhibit different sentence structures. However, the fact that the BT-from-WFW model achieves the highest BLEU and chrF scores on both test sets, albeit slightly, suggests that there is still something to be gained from training with WFW-BT data on the first run, especially considering that all the models have already seen the entire monolingual data during mBart pre-training.

Experiment	WMT Test		BBC Test	
	BLEU	chrF	BLEU	chrF
Baseline	8.3	31.2	9.0	34.2
WFW-Bivec	8.4	31.2	9.4	35.0
WFW-MUSE	8.2	30.7	9.2	34.6
BT	9.1	31.9	12.3	37.7
BT-from-WFW	**9.4**	**32.4**	**12.5**	**38.5**

Table 5: BLEU and chrF for the NMT models described in Section 4.6 in the En-Ps direction.

On the other hand, Table 6 shows the metrics for the Ps-En translation direction. In comparison to Table 5 the metrics are significantly higher across the board. This is likely an upshot of the mBart pre-training objective as the Ps-En direction uses the entire English paracrawl corpus, resulting in a stronger decoder performance. WFW-Bivec again outperforms both the Baseline and WFW-MUSE although the latter is closer in all metrics and both have a chrF of 42.7 on the BBC test set. Back-translation also leads to a significant increase in performance. However, BT-from-WFW shows small but consistent improvements for all metrics.

Experiment	WMT Test		BBC Test	
	BLEU	chrF	BLEU	chrF
Baseline	12.1	37.4	14.8	42.1
WFW-Bivec	12.2	37.7	15.0	42.7
WFW-MUSE	12.0	37.5	14.6	42.7
BT	13.6	39.7	18.8	47.9
BT-from-WFW	**13.8**	**39.9**	**19.0**	**48.1**

Table 6: BLEU and chrF for the NMT models described in Section 4.6 in the Ps-En direction.

The micro-averaged F1 scores for OOV words given in Figure 1 are on the whole low. Low recalls drive the low F1 scores at all frequencies. For all models, WFW-Bivec results in a higher F1 than the MUSE-based method in the En→Ps direction, supporting the fact that Bivec Combined results in better WFW translations. It is also evident that the F1 score is higher at all

frequencies for the BT-from-WFW model that uses back-translations from the corresponding WFW-Bivec model compared to just using back-translations from the Baseline model for the En-Ps NMT models.

However, in the Ps-En direction, the situation is less clear, where BT outperforms BT-from-WFW not only at the frequencies of 15 and 25 but also for OOV words. The difference at the other frequencies, while favouring BT-from-WFW, is slight, and hence it seems likely that the improved BT-from-WFW metrics presented in Table 6 are not due to the models learning better translations of OOV or rare words.

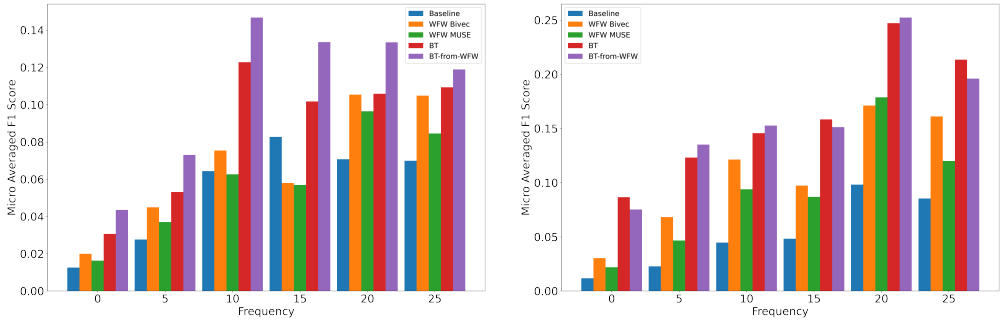

Figure 1: Micro averaged F1 scores for OOV and rare words in BBC est set at frequencies. **Left:** En-Ps **Right:** Ps-En.

Across the board, the increase in the F1 score of the WFW based corpora is significantly smaller than that of adding back-translations. This is even the case for OOV words, meaning that back-translation learns the correct translation of more OOV words. Further confirmation of this can be seen when looking at the recall of OOVs for the WFW-Bivec models, which are 0.021 for En→Ps and 0.031 for Ps→En. Such low recalls illustrate that models are learning very few OOV words from the WFW pseudo-parallel corpus.

6 Conclusion

The BWE evaluation results show that MUSE correctly translates more OOV words than the proposed Combined Bivec approach, where more weight is given to sentence-level accuracy results as they cover a higher proportion of OOV words. However, when viewed in the context of NMT, it appears that the WFW back-translations using Bivec lead to more OOV and rare words being correctly predicted. For OOV words we hypothesise that this is due to the WFW-BT model being able to leverage the higher overall quality of the Bivec Combined back-translations to predict more OOV words correctly. Specifically, this means that Bivec Combined results in a higher proportion of context words of the OOV word being translated correctly.

Regarding NMT, the results demonstrate that incorporating word-level translations benefits the model even when using back-translation when the low-resource language is on the target side. However, the results are less conclusive when the high-resource language is the target language. The low recall for all OOV words for the NMT task suggests that even when the dictionary contains accurate translations, it is difficult for these to transfer into correct model predictions. As a result, it seems that incorporating word-level translations from the monolingual data can benefit the model. It may be that for languages that exhibit a different sentence structure, WFW back-translation is not the best methodology for incorporating the OOVs and rare words. Instead, an approach of inserting them into back-translations or existing parallel data may be more appropriate to ensure a higher degree of fluency in the synthetic sentences.

Acknowledgement

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 825299 (GoURMET), the European Research Council (ERC StG BroadSem 678254; ERC CoG TransModal 681760) and funding by the UK Engineering and Physical Sciences Research Council (EPSRC) fellowship grant EP/S001271/1 (MTStretch). Jonas Waldendorf is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh.

References

Artetxe, M., Labaka, G., and Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., and Agirre, E. (2020). A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Birch, A., Haddow, B., Valerio Miceli Barone, A., Helcl, J., Waldendorf, J., Sánchez Martínez, F., Forcada, M., Sánchez Cartagena, V., Pérez-Ortiz, J. A., Esplà-Gomis, M., Aziz, W., Murady, L., Sariisik, S., van der Kreeft, P., and Macquarrie, K. (2021). Surprise language challenge: Developing a neural machine translation system between Pashto and English in two months. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 92–102, Virtual. Association for Machine Translation in the Americas.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chronopoulou, A., Stojanovski, D., and Fraser, A. (2021). Improving the lexical ability of pretrained language models for unsupervised neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Online. Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dyer, C., Chahuneau, V., and Smith, N. A. (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.

Eder, T., Hangya, V., and Fraser, A. (2020). Anchor-based Bilingual Word Embeddings for Low-Resource Languages. *arXiv:2010.12627 [cs]*. arXiv: 2010.12627.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Hu, J., Xia, M., Neubig, G., and Carbonell, J. (2019). Domain Adaptation of Neural Machine Translation by Lexicon Induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Huck, M., Hangya, V., and Fraser, A. (2019). Better OOV Translation with Bilingual Terminology Mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5809–5815, Florence, Italy. Association for Computational Linguistics.

Junczys-Dowmunt, M., Heafield, K., Hoang, H., Grundkiewicz, R., and Aue, A. (2018). Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., and Guzmán, F. (2020). Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., and Li, L. (2020). Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168 [cs]*. arXiv: 1309.4168.

Ormazabal, A., Artetxe, M., Labaka, G., Soroa, A., and Agirre, E. (2019). Analyzing the Limitations of Cross-lingual Word Embedding Mappings. *arXiv:1906.05407 [cs]*. arXiv: 1906.05407.

Peng, W., Huang, C., Li, T., Chen, Y., and Liu, Q. (2020). Dictionary-based Data Augmentation for Cross-Domain Neural Machine Translation. *arXiv:2004.02577 [cs]*. arXiv: 2004.02577.

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

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

Shen, J., Chen, P.-J., Le, M., He, J., Gu, J., Ott, M., Auli, M., and Ranzato, M. (2021). The source-target domain mismatch problem in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1519–1533, Online. Association for Computational Linguistics.

Søgaard, A., Ruder, S., and Vulić, I. (2018). On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Xu, R., Zhi, Z., Cao, J., Wang, M., and Li, L. (2020). Volctrans parallel corpus filtering system for WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 985–990, Online. Association for Computational Linguistics.