Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings

Phillip Keung[•] **Julian Salazar**[•] **Yichao Lu**[•] **Noah A. Smith**[†‡]

[•]Amazon [†]University of Washington [‡]Allen Institute for AI
{keung,julsal,yichaolu}@amazon.com nasmith@cs.washington.edu

Abstract

We describe an unsupervised method to create *pseudo-parallel corpora* for machine translation (MT) from unaligned text. We use multilingual BERT to create source and target sentence embeddings for nearest-neighbor search and adapt the model via self-training. We validate our technique by extracting parallel sentence pairs on the BUCC 2017 bitext mining task and observe up to a 24.5 point increase (absolute) in F_1 scores over previous unsupervised methods. We then improve an XLM-based unsupervised neural MT system pre-trained on Wikipedia by supplementing it with pseudo-parallel text mined from the same corpus, boosting unsupervised translation performance by up to 3.5 BLEU on the WMT'14 French-English and WMT'16 German-English tasks and outperforming the previous state-of-the-art. Finally, we enrich the IWSLT'15 English-Vietnamese corpus with pseudo-parallel Wikipedia sentence pairs, yielding a 1.2 BLEU improvement on the low-resource MT task. We demonstrate that unsupervised bitext mining is an effective way of augmenting MT datasets and complements existing techniques like initializing with pre-trained contextual embeddings.

1 Introduction

Large corpora of parallel sentences are prerequisites for training models across a diverse set of applications, such as neural machine translation (NMT; Bahdanau et al., 2015), paraphrase generation (Bannard and Callison-Burch, 2005), and aligned multilingual sentence embeddings (Artetxe and Schwenk, 2019b). Systems that extract parallel corpora typically rely on various cross-lingual resources (e.g., bilingual lexicons, parallel corpora), but recent work has shown that unsupervised parallel sentence mining (Hangya et al., 2018) and unsupervised NMT (Artetxe et al., 2018; Lample et al., 2018a) produce surprisingly good results.[1]

Existing approaches to unsupervised parallel sentence (or *bitext*) mining start from bilingual word embeddings (BWEs) learned via an unsupervised, adversarial approach (Lample et al., 2018b). Hangya et al. (2018) created sentence representations by mean-pooling BWEs over content words. To disambiguate semantically similar but non-parallel sentences, Hangya and Fraser (2019) additionally proposed parallel segment detection by searching for paired substrings with high similarity scores per word. However, using word embeddings to generate sentence embeddings ignores sentential context, which may degrade bitext retrieval performance.

We describe a new unsupervised bitext mining approach based on contextual embeddings. We create sentence embeddings by mean-pooling the outputs of multilingual BERT (mBERT; Devlin et al., 2019), which is pre-trained on unaligned Wikipedia sentences across 104 languages. For a pair of source and target languages, we find candidate translations by using nearest-neighbor search with margin-based similarity scores between pairs of mBERT-embedded source and target sentences. We bootstrap a dataset of positive and negative sentence pairs from these initial neighborhoods of candidates, then self-train mBERT on its own outputs. A final retrieval step gives a corpus of *pseudo-parallel* sentence pairs, which we expect to be a mix of actual translations and semantically related non-translations.

[1]By *unsupervised*, we mean that no cross-lingual resources like parallel text or bilingual lexicons are used. Unsupervised techniques have been used to bootstrap MT systems for low-resource languages like Khmer and Burmese (Marie et al., 2019).

Transactions of the Association for Computational Linguistics, vol. 8, pp. 828–841, 2020. https://doi.org/10.1162/tacl_a_00348
Action Editor: Collin Cherry. Submission batch: 4/2020; Revision batch: 8/2020; Published 12/2020.

We apply our technique on the BUCC 2017 parallel sentence mining task (Zweigenbaum et al., 2017). We achieve state-of-the-art F_1 scores on unsupervised bitext mining, with an improvement of up to 24.5 points (absolute) on published results (Hangya and Fraser, 2019). Other work (e.g., Libovický et al., 2019) has shown that retrieval performance varies substantially with the layer of mBERT used to generate sentence representations; using the optimal mBERT layer yields an improvement as large as 44.9 points.

Furthermore, our pseudo-parallel text improves unsupervised NMT (UNMT) performance. We build upon the UNMT framework of Lample et al. (2018c) and XLM (Lample and Conneau, 2019) by incorporating our pseudo-parallel text (also derived from Wikipedia) at training time. This boosts performance on WMT'14 En-Fr and WMT'16 En-De by up to 3.5 BLEU over the XLM baseline, outperforming the state-of-the-art on unsupervised NMT (Song et al., 2019).

Finally, we demonstrate the practical value of unsupervised bitext mining in the low-resource setting. We augment the English-Vietnamese corpus (133k pairs) from the IWSLT'15 translation task (Cettolo et al., 2015) with our pseudo-bitext from Wikipedia (400k pairs), and observe a 1.2 BLEU increase over the best published model (Nguyen and Salazar, 2019). When we reduced the amount of parallel and monolingual Vietnamese data by a factor of ten (13.3k pairs), the model trained with pseudo-bitext performed 7 BLEU points better than a model trained on the reduced parallel text alone.

2 Our Approach

Our aim is to create a bilingual sentence embedding space where, for each source sentence embedding, a sufficiently close nearest neighbor among the target sentence embeddings is its translation. By aligning source and target sentence embeddings in this way, we can extract sentence pairs to create new parallel corpora. Artetxe and Schwenk (2019a) construct this space by training a joint encoder-decoder MT model over multiple language pairs and using the resulting encoder to generate sentence embeddings. A margin-based similarity score is then computed between embeddings for retrieval (Section 2.2). However, this approach requires large parallel corpora to train the encoder-decoder model in the first place.

We investigate whether contextualized sentence embeddings created with unaligned text are useful for *unsupervised bitext retrieval*. Previous work explored the use of multilingual sentence encoders taken from machine translation models (e.g., Artetxe and Schwenk, 2019b; Lu et al., 2018) for zero-shot cross-lingual transfer. Our work is motivated by recent success in tasks like zero-shot text classification and named entity recognition (e.g., Keung et al., 2019; Mulcaire et al., 2019) with multilingual contextual embeddings, which exhibit cross-lingual properties despite being trained without parallel sentences.

We illustrate our method in Figure 1. We first retrieve the candidate translation pairs:

- Each source and target language sentence is converted into an embedding vector with mBERT via mean-pooling.
- Margin-based scores are computed for each sentence pair using the k nearest neighbors of the source and target sentences (Sec. 2.2).
- Each source sentence is paired with its nearest neighbor in the target language based on this score.
- We select a threshold score that keeps some top percentage of pairs (Sec. 2.2).
- Rule-based filters are applied to further remove mismatched sentence pairs (Sec. 2.3).

The remaining candidate pairs are used to bootstrap a dataset for self-training mBERT as follows:

- Each candidate pair (a source sentence and its closest nearest neighbor above the threshold) is taken as a positive example.
- This source sentence is also paired with its next $k - 1$ neighbors to give hard negative examples (we compare this with random negative samples in Sec. 3.3).
- We finetune mBERT to produce sentence embeddings that discriminate between positive and negative pairs (Sec. 2.4).

After self-training, the finetuned mBERT model is used to generate new sentence embeddings. Parallel sentences should be closer to each other in this new embedding space, which improves retrieval performance.

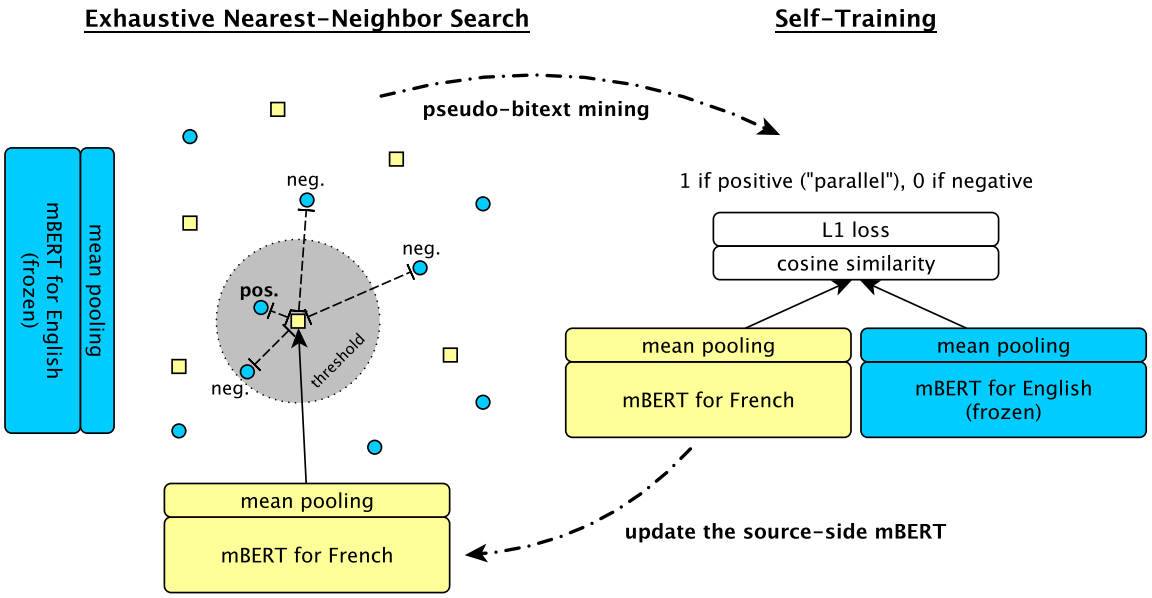

Figure 1: Our self-training scheme. **Left:** We index sentences using our two encoders. For each source sentence, we retrieve k nearest-neighbor target sentences per the margin criterion (Eq. 1), depicted here for $k = 4$. If the nearest neighbor is within a threshold, it is treated with the source sentence as a positive pair, and the remaining $k - 1$ are treated with the source sentence as negative pairs. **Right:** We refine one of the encoders such that the cosine similarity of the two embeddings is maximized on positive pairs and minimized on negative pairs.

2.1 Sentence Embeddings and Nearest-neighbor Search

We use mBERT (Devlin et al., 2019) to create sentence embeddings for both languages by mean-pooling the representations from the final layer. We use FAISS (Johnson et al., 2017) to perform exact nearest-neighbor search on the embeddings. We compare every sentence in the source language to every sentence in the target language; we do not use links between Wikipedia articles or other metadata to reduce the size of the search space. In our experiments, we retrieve the $k = 4$ closest target sentences for each source sentence; the source language is always non-English, while the target language is always English.

2.2 Margin-based Score

We compute a margin-based similarity score between each source sentence and its k nearest target neighbors. Following Artetxe and Schwenk (2019a), we use the *ratio* margin score, which calibrates the cosine similarity by dividing it by the average cosine distance of each embedding's k nearest neighbors:

$$\text{margin}(x, y) = \frac{\cos(x, y)}{\sum_{z \in \text{NN}_k^{\text{tgt}}(x)} \frac{\cos(x,z)}{2k} + \sum_{z \in \text{NN}_k^{\text{src}}(y)} \frac{\cos(y,z)}{2k}}. \quad (1)$$

We remove the sentence pairs with margin scores below some pre-selected threshold. For BUCC, we do not have development data for tuning the threshold hyperparameter, so we simply use the prior probability. For example, the creators of the dataset estimate that ∼2% of De sentences have an En translation, so we choose a score threshold such that we retrieve ∼2% of the pairs. We set the threshold in the same way for the other BUCC pairs. For UNMT with Wikipedia bitext mining, we set the threshold such that we always retrieve 2.5 million sentence pairs for each language pair.

2.3 Rule-based Filtering

We also apply two simple filtering steps before finalizing the candidate pairs list:

- **Digit filtering**: Sentence pairs that are translations of each other must have digit sequences that match exactly.[2]
- **Edit distance**: Sentences from English Wikipedia sometimes appear in non-English pages and vice versa. We remove sentence pairs where the content of the source and

[2] In Python, `set(re.findall("[0-9]+",sent1)) == set(re.findall("[0-9]+",sent2))`.

target share substantial overlap (i.e., the character-level edit distance is $\leq$50%).

2.4 Self-training

We devise an unsupervised self-training technique to improve mBERT for bitext retrieval using mBERT's own outputs. For each source sentence, if the nearest target sentence is within the threshold and not filtered out, the pair is treated as a positive sentence. We then keep the next $k-1$ nearest neighbors as negative sentences. Altogether, these give us a training set of examples which are labeled as positive or negative pairs.

We train mBERT to discriminate between positive and negative sentence pairs as a binary classification task. We distinguish the mBERT encoders for the source and target languages as f_{src}, f_{tgt} respectively. Our training objective is

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}; \Theta_{\text{src}}) = \left| \frac{f_{\text{src}}(\boldsymbol{X}; \Theta_{\text{src}})^\top f_{\text{tgt}}(\boldsymbol{Y})}{\|f_{\text{src}}(\boldsymbol{X}; \Theta_{\text{src}})\| \|f_{\text{tgt}}(\boldsymbol{Y})\|} - \text{Par}(\boldsymbol{X}, \boldsymbol{Y}) \right|, \quad (2)$$

where $f_{\text{src}}(\boldsymbol{X})$ and $f_{\text{tgt}}(\boldsymbol{Y})$ are the mean-pooled representations of the source sentence $\boldsymbol{X}$ and target sentence $\boldsymbol{Y}$, and where $\text{Par}(\boldsymbol{X}, \boldsymbol{Y})$ is 1 if $\boldsymbol{X}, \boldsymbol{Y}$ are parallel and 0 otherwise. This loss encourages the cosine similarity between the source and target embeddings to increase for positive pairs and decrease otherwise. The process is depicted in Figure 1.

Note that we only finetune f_{src} (parameters Θ_{src}) and we hold f_{tgt} fixed. If both f_{src} and f_{tgt} are updated, then the training process collapses to a trivial solution, since the model will map all pseudo-parallel pairs to one representation and all non-parallel pairs to another. We hold f_{tgt} fixed, which forces f_{src} to align its outputs to the target (in our experiments, always English) mBERT embeddings.

After finetuning, we use the updated f_{src} to generate new non-English sentence embeddings. We then repeat the retrieval process with FAISS, yielding a final set of pseudo-parallel pairs after thresholding and filtering.

3 Unsupervised Bitext Mining

We apply our method to the BUCC 2017 shared task, ''Spotting Parallel Sentences in Comparable Corpora'' (Zweigenbaum et al., 2017). The task involves retrieving parallel sentences from monolingual corpora derived from Wikipedia. Parallel sentences were inserted into the corpora in a contextually appropriate manner by the task organizers. The shared task assessed retrieval systems for precision, recall, and F_1-score on four language pairs: De-En, Fr-En, Ru-En, and Zh-En. Prior work on unsupervised bitext mining has generally studied the European language pairs to avoid dealing with Chinese word segmentation (Hangya et al., 2018; Hangya and Fraser, 2019).

3.1 Setup

For each BUCC language pair, we take the corresponding source and target monolingual corpus, which have been pre-split into *training*, *sample*, and *test* sets at a ratio of 49%–2%–49%. The identity of the parallel sentence pairs for the test set were not publicly released, and are only available for the training set. Following the convention established in Hangya and Fraser (2019) and Artetxe and Schwenk (2019a), we use the *test* portion for unsupervised system development and evaluate on the *training* portion.

We use the reference FAISS implementation[3] for nearest-neighbor search. We used the GluonNLP toolkit (Guo et al., 2020) with pre-trained mBERT weights[4] for inference and self-training. We compute the margin similarity score in Eq. 1 with $k = 4$ nearest neighbors. We set a threshold on the score such that we retrieve the prior proportion (e.g., $\sim$2%) of parallel pairs in each language.

We then finetune mBERT via self-training. We take minibatches of 100 sentence pairs. We use the Adam optimizer with a constant learning rate of 0.00001 for 2 epochs. To avoid noisy translations, we finetune on the top 50% of the highest-scoring pairs from the retrieved bitext (e.g., if the prior proportion is 2%, then we would use the top 1% of sentence pairs for self-training).

We considered performing more than one round of self-training but found it was not helpful for the BUCC task. BUCC has very few parallel pairs (e.g., 9,000 pairs for Fr-En) per language and thus few positive pairs for our unsupervised method to find. The size of the self-training corpus is limited by the proportion of parallel sentences, and mBERT rapidly overfits to small datasets.

[3] https://github.com/facebookresearch/faiss.

[4] https://github.com/google-research/bert/blob/master/multilingual.md.

Method	De-En	Fr-En	Ru-En	Zh-En
Hangya and Fraser (2019)				
avg.	30.96	44.81	19.80	–
align-static	42.81	42.21	24.53	–
align-dyn.	43.35	43.44	24.97	–
Our method				
mBERT (final layer)	42.1	45.8	36.9	35.8
+ digit filtering (DF)	47.0	49.3	41.2	38.0
+ edit distance (ED)	47.0	49.3	41.2	38.0
+ self-training (ST)	**60.6**	**60.2**	**49.5**	**45.7**
mBERT (layer 8)	67.0	65.3	59.3	53.3
+ DF, ED, ST	**74.9**	**73.0**	**69.9**	**60.1**

Table 1: F_1 scores for unsupervised bitext retrieval on BUCC 2017. Results with mBERT are from our method (Sec. 2) using the final (12th) layer. We also include results for the 8th layer (e.g., Libovický et al., 2019), but do not consider this part of the unsupervised setting as we would not have known a priori which layer was best to use.

Language pair	Parallel sentence pair
De-En	Beide Elemente des amerikanischen Traums haben heute einen Teil ihrer Anziehungskraft verloren. Both elements of the American dream have now lost something of their appeal.
Fr-En	L'Allemagne à elle seule s'attend à recevoir pas moins d'un million de demandeurs d'asile cette année. Germany alone expects as many as a million asylum-seekers this year.
Ru-En	Однако по решению Берлинского конгресса в 1881 году к территории Греции присоединилась Фессалия и часть Эпира. Nevertheless, in 1881, Thessaly and small parts of Epirus were ceded to Greece as part of the Treaty of Berlin.
Zh-En	在如今这个奇怪的新世界里，现代和前现代相互依存。 In the strange new world of today, the modern and the pre-modern depend on each other.

Table 2: Examples of parallel sentences that were extracted by our method on the BUCC 2017 shared task.

3.2 Results

We provide a few examples of the bitext we retrieved in Table 2. The examples were chosen from the high-scoring pairs and verified to be correct translations.

Our retrieval results are in Table 1. We compare our results with strictly unsupervised techniques, which do not use bilingual lexicons, parallel text, or other cross-lingual resources. Using mBERT as-is with the margin-based score works reasonably well, giving F_1 scores in the range of 35.8 to 45.8, which is competitive with the previous state-of-the-art for some pairs, and outperforming by 12 points in the case of Ru-En. Furthermore, applying simple rule-based filters (Sec. 2.3) on the candidate translation pairs adds a few more points, although the edit distance filter has a negligible effect when compared with the digit filter.

Method	De-En	Fr-En	Ru-En	Zh-En
mBERT w/o ST	47.0	49.3	41.2	38.0
w/ ST (random)	57.7	55.7	48.1	45.2
w/ ST (hard)	**60.6**	**60.2**	**49.5**	**45.7**

Table 3: F_1 scores for bitext retrieval on BUCC 2017 using random sentences as negative samples instead of nearest neighbors.

We see that finetuning mBERT on its own chosen sentence pairs (i.e., unsupervised self-training) yields significant improvements, adding another 8 to 14 points to the F_1 score on top of filtering. In all, these F_1 scores represent a 34% to 98% relative improvement over existing techniques in unsupervised parallel sentence extraction for these language pairs.

Libovický et al. (2019) explored bitext mining with mBERT in the supervised context and found that retrieval performance significantly varies with the mBERT layer used to create sentence embeddings. In particular, they found layer 8 embeddings gave the highest precision-at-1. We also observe an improvement (Table 1) in unsupervised retrieval of another 13 to 20 points by using the 8th layer instead of the default final layer (12th). We include these results but do not consider them unsupervised, as we would not know *a priori* which layer was best to use.

3.3 Choosing Negative Sentence Pairs

Other authors (e.g., Guo et al., 2018) have noted that the choice of negative examples has a considerable impact on metric learning. Specifically, using negative examples which are difficult to distinguish from the positive nearest neighbor is often beneficial for performance. We examine the impact of taking random sentences instead of the remaining $k-1$ nearest neighbors as the negatives during self-training.

Our results are in Table 3. While self-training with random negatives still greatly improves the untuned baseline, the use of hard negative examples mined from the k-nearest neighborhood can make a significant difference to the final F_1 score.

4 Bitext for Neural Machine Translation

A major application of bitext mining is to create new corpora for machine translation. We conduct an extrinsic evaluation of our unsupervised bitext mining approach on *unsupervised* (WMT'14 French-English, WMT'16 German-English) and *low-resource* (IWSLT'15 English-Vietnamese) translation tasks.

We perform large-scale unsupervised bitext extraction on the October 2019 Wikipedia dumps in various languages. We use `wikifil.pl`[5] to extract paragraphs from Wikipedia and remove markup. We then use the `syntok`[6] package for sentence segmentation. Finally, we reduce the size of the corpus by removing sentences that aren't part of the body of Wikipedia pages. Sentences that contain `*`, `=`, `//`, `::`, `#`, `www`, `(talk)`, or the pattern `[0-9]{2}:[0-9]{2}` are filtered out.

We index, retrieve, and filter candidate sentence pairs with the procedure in Sec. 3. Unlike BUCC, the Wikipedia dataset does not fit in GPU memory. The processed corpus is quite large, with 133 million, 67 million, 36 million, and 6 million sentences in English, German, French, and Vietnamese respectively. We therefore shard the dataset into chunks of 32,768 sentences and perform nearest-neighbor comparisons in chunks for each language pair. We use a simple map-reduce algorithm to merge the intermediate results back together.

We follow the approach outlined in Sec. 2 for Wikipedia bitext mining. For each source sentence, we retrieve the four nearest target neighbors across the millions of sentences that we extracted from Wikipedia and compute the margin-based scores for each pair.

4.1 Unsupervised NMT

We show that our pseudo-parallel text can complement existing techniques for unsupervised translation (Artetxe et al., 2018; Lample et al., 2018c). In line with existing work on UNMT, we evaluate our approach on the WMT'14 Fr-En and WMT'16 De-En test sets.

Our UNMT experiments build upon the reference implementation[7] of XLM (Lample and Conneau, 2019). The UNMT model is trained by alternating between two steps: a denoising autoencoder step and a backtranslation step (refer to Lample et al., 2018c for more details). The backtranslation step generates pseudo-parallel

[5] `https://github.com/facebookresearch/fastText/blob/master/wikifil.pl`.

[6] `https://github.com/fnl/syntok`.

[7] `https://github.com/facebookresearch/xlm`.

Reference	Architecture	Pre-training	En-De	De-En	En-Fr	Fr-En
Artetxe et al. (2018)	2-layer RNN		6.89	10.16	15.13	15.56
Lample et al. (2018a)	3-layer RNN		9.75	13.33	15.05	14.31
Yang et al. (2018)	4-layer Transformer		10.86	14.62	16.97	15.58
Lample et al. (2018c)	4-layer Transformer		17.16	21.00	25.14	24.18
Song et al. (2019)	6-layer Transformer	MASS	28.3	35.2	37.5	34.9
XLM Baselines						
Lample and Conneau (2019)	6-layer Transformer	XLM	–	–	33.4	33.3
Song et al. (2019)	6-layer Transformer	XLM	27.0	34.3	33.4	33.3
XLM reference implementation	6-layer Transformer	XLM	–	–	36.6	34.0
Maximum performance across baselines	6-layer Transformer	XLM	27.0	34.3	36.6	34.0
Ours						
Our XLM baseline	6-layer Transformer	XLM	27.7	34.5	36.7	34.5
w/ pseudo-parallel text before ST	6-layer Transformer	XLM	30.4	36.3	39.7	35.9
w/ pseudo-parallel text after ST	6-layer Transformer	XLM	**30.7**	**37.3**	**40.2**	**36.9**

Table 4: BLEU scores for unsupervised NMT performance on WMT'14 English-French and WMT'16 English-German test sets. All methods only use unaligned Wikipedia corpora for pre-training and/or bitext mining. 'ST' refers to self-training.

training data, and we incorporate our bitext during UNMT training in the same way, as another set of pseudo-parallel sentences. We also use the same initialization as Lample and Conneau (2019), where the UNMT models have encoders and decoders that are initialized with contextual embeddings trained on the source and target language Wikipedia corpora with the masked language model (MLM) objective; no parallel data is used.

We performed the exhaustive (Fr Wiki)-(En Wiki) and (De Wiki)-(En Wiki) nearest-neighbor comparison on eight V100 GPUs, which requires 3 to 4 days to complete per language pair. We retained the top 2.5 million pseudo-parallel Fr-En and De-En sentence pairs after mining.

4.2 Results

Our results are in Table 4. The addition of mined bitext consistently increases the BLEU score in both directions for WMT'14 Fr-En and WMT'16 De-En. Much of the existing work on improving UNMT focuses on improved initialization with contextual embeddings like XLM or MASS (Song et al., 2019). These embeddings were already pre-trained on Wikipedia data, so it is surprising that adding our pseudo-parallel Wikipedia sentences leads to a 2 to 3 BLEU improvement. In other words, our approach is complementary to pre-trained initialization techniques.

Previously (in Table 1), we saw that self-training improved the F_1 score for BUCC bitext retrieval. The improvement in bitext quality carries over to UNMT, and providing better pseudo-parallel text yields a consistent improvement for all translation directions.

Our results are state-of-the-art in UNMT, but they should be interpreted relative to the strength of our XLM baseline. We are building on top of the XLM initialization, and the effectiveness of the initialization (and the various hyperparameters used during training and decoding) affects the strength of our final results. For example, we adjusted the beam width on our XLM baselines to attain BLEU scores which are similar to what others have published. One can apply our method to MASS, which performs better than XLM on UNMT, but we chose to report results on XLM because it has been validated on a wider range of tasks and languages.

We also trained a standard 6-layer transformer encoder-decoder model directly on the pseudo-parallel text. We used the standard implementation in Sockeye (Hieber et al., 2018) as-is, and trained models for French and German on 2.5 million Wikipedia sentence pairs. We withheld 10k pseudo-parallel pairs per language pair to serve as a development set. We achieved BLEU scores of 20.8, 21.1, 28.2, and 28.0 on En-De, De-En, En-Fr, and Fr-En respectively. BLEU scores were computed with SacreBLEU (Post, 2018).

This compares favorably with the best UNMT results in Lample et al. (2018c), while avoiding the use of parallel development data altogether.

4.3 Low-resource NMT

French and German are high-resource languages and are linguistically close to English. We therefore evaluate our mined bitext on a low-resource, linguistically distant language pair. The IWSLT'15 English-Vietnamese MT task (Cettolo et al., 2015) provides 133k sentence pairs derived from translated TED talks transcripts and is a common benchmark for low-resource MT. We take supervised training data from the IWSLT task and augment it with different amounts of pseudo-parallel text mined from English and Vietnamese Wikipedia. Furthermore, we construct a very low-resource setting by downsampling the parallel text and monolingual Vietnamese Wikipedia text by a factor of ten (13.3k sentence pairs).

We use the reference implementation[8] for the state-of-the-art model (Nguyen and Salazar, 2019), which is a highly regularized 6+6-layer transformer with pre-norm residual connections, scale normalization, and normalized word embeddings. We use the same hyperparameters (except for the dropout rate) but train on our augmented datasets. To mitigate domain shift, we finetune the best checkpoint for 75k more steps using only the IWSLT training data, in the spirit of ''trivial'' transfer learning for low-resource NMT (Kocmi and Bojar, 2018).

In Table 5, we show BLEU scores as more pseudo-parallel text is included during training. As in previous works on En-Vi (cf. Luong and Manning, 2015), we use tst2012 (1,553 pairs) and tst2013 (1,268 pairs) as our development and test sets respectively, we tokenize all data with Moses, and we report tokenized BLEU via `multi-bleu.perl`. The BLEU score increases monotonically with the size of the pseudo-parallel corpus and exceeds the state-of-the-art system's BLEU by 1.2 points. This result is consistent with improvements observed with other types of monolingual data augmentation like pre-trained UNMT initialization, various forms of back-translation (Hoang et al., 2018; Zhou and Keung, 2020), and cross-view training (CVT; Clark et al., 2018):

[8] `https://github.com/tnq177/transformers_without_tears`.

	En-Vi
Luong and Manning (2015)	26.4
Clark et al. (2018)	28.9
Clark et al. (2018), with CVT	29.6
Xu et al. (2019)	31.4
Nguyen and Salazar (2019)	32.8 (28.8)
+ top 100k mined pairs	33.2 (29.5)
+ top 200k mined pairs	33.9 (29.8)
+ top 300k mined pairs	**34.0 (30.0)**
+ top 400k mined pairs	34.1 (29.9)

Table 5: Tokenized BLEU scores on tst2013 for the low-resource IWSLT'15 English-Vietnamese translation task using bitext mined with our method. Added pairs are sorted by their score. Development scores on tst2012 in parentheses.

We describe our hyperparameter tuning and infrastructure following Dodge et al. (2019). The translation sections of this work mostly used default parameters, but we did tune the dropout rate (at 0.2 and 0.3) for each amount of mined bitext for the supervised En-Vi task (at 100k, 200k, 300k, and 400k sentence pairs). We include development scores for our best models; dropout of 0.3 did best for 0k and 100k, while 0.2 did best otherwise. Training takes less than a day on one V100 GPU.

To simulate a very low-resource task, we use one-tenth of the training data by downsampling the IWSLT En-Vi train set to 13.3k sentence pairs. Furthermore, we mine bitext from one-tenth of the monolingual Wiki Vi text and extract proportionately fewer sentence pairs (i.e., 10k, 20k, 30k, and 40k pairs). We use the implementation and hyperparameters for the regularized 4+4-layer transformer used by Nguyen and Salazar (2019) in a similar setting. We tune the dropout rate (0.2, 0.3, 0.4) to maximize development performance; 0.4 was best for 0k, 0.3 for 10k and 20k, and 0.2 for 30k and 40k. In Table 6, we see larger improvements in BLEU (4+ points) for the same relative increases in mined data (as compared to Table 5). In both cases, the rate of improvement tapers off as the quality and relative quantity of mined pairs degrades at each increase.

4.4 UNMT Ablation Study: Pre-training and Bitext Mining Corpora

In Sec. 4.2, we mined bitext from the October 2019 Wikipedia snapshot whereas the pre-trained

	En-Vi, one-tenth
13.3k pairs (from 133k original)	20.7 (19.5)
+ top 10k mined pairs	25.0 (22.9)
+ top 20k mined pairs	26.7 (24.1)
+ top 30k mined pairs	27.3 (24.5)
+ top 40k mined pairs	**27.7 (24.7)**

Table 6: Tokenized BLEU scores (tst2013), where the bitext was mined from one-tenth of the monolingual Vietnamese data. Development scores on tst2012 in parentheses.

XLM embeddings were created prior to January 2019. Hence, it is possible that the UNMT BLEU increase would be smaller if the bitext were mined from the same corpus used for pre-training. We ran an ablation study to show the effect (or lack thereof) of the overlap between the pre-training and pseudo-parallel corpora.

For the En-Vi language pair, we used 5 million English and 5 million Vietnamese Wiki sentences to pre-train the XLM model. We only use text from the October 2019 Wiki snapshot. We mined 300k pseudo-parallel sentence pairs using our approach (Sec. 2) from the same Wiki snapshot. We created two datasets for XLM pre-training: a 10 million-sentence corpus that is disjoint from the 600k sentences of the mined bitext, and a 10 million-sentence corpus that contains all 600k sentences of the bitext. In Table 7, we show the BLEU increase on the IWSLT En-Vi task with and without using the mined bitext as parallel data, using each of the two XLM models as the initialization.

The benefit of using pseudo-parallel text is very clear; even if the pre-trained XLM model saw the pseudo-parallel sentences during pre-training, using mined bitext still significantly improves UNMT performance (23.1 vs. 28.3 BLEU). In addition, the baseline UNMT performance without the mined bitext is similar between the two XLM initializations (23.1 vs. 23.2 BLEU), which suggests that removing some of the parallel text present during pre-training does not have a major effect on UNMT.

Finally, we trained a standard encoder-decoder model on the 300k pseudo-parallel pairs only, using the same Sockeye recipe in Sec. 4.2. This yielded a BLEU score of 27.5 on En-Vi, which is lower than the best XLM-based result (i.e., 28.9), which suggests that the XLM initialization improves unsupervised NMT. A similar outcome was also reported in Lample and Conneau (2019).

	w/o PP as bitext	w/ PP as bitext
XLM excl. PP text	23.2	28.9
XLM incl. PP text	23.1	28.3

Table 7: Tokenized UNMT BLEU scores on IWSLT'15 English-Vietnamese (tst2013) with XLM initialization. We mined 300k pseudo-parallel (PP) sentence pairs from En and Vi Wikipedia (Oct. 2019). We created two XLM models, with the pre-training corpus including or excluding the PP pairs. We compare their downstream UNMT performance with and without PP pairs as ''bitext'' during UNMT training.

5 Related Work

5.1 Parallel Sentence Mining

Approaches to parallel sentence (or bitext) mining have been historically driven by the data requirements of statistical machine translation. Some of the earliest work in mining the Web for large-scale parallel corpora can be found in Resnik (1998) and Resnik and Smith (2003). Recent interest in the field is reflected by new shared tasks on parallel extraction and filtering (Zweigenbaum et al., 2017; Koehn et al., 2018) and the creation of massively multilingual parallel corpora mined from the Web, like WikiMatrix (Schwenk et al., 2019a) and CCMatrix (Schwenk et al., 2019b).

Existing parallel corpora have been exploited in many ways to create sentence representations for supervised bitext mining. One approach involves a joint encoder with a shared wordpiece vocabulary, trained as part of multiple encoder-decoder translation models on parallel corpora (Schwenk, 2018). Artetxe and Schwenk (2019b) apply this approach at scale, and shared a single encoder and joint vocabulary across 93 languages. Another approach uses negative sampling to align the encoders' sentence representations for nearest-neighbor retrieval (Grégoire and Langlais, 2018; Guo et al., 2018).

However, these approaches require training with initial parallel corpora. In contrast, Hangya et al. (2018) and Hangya and Fraser (2019) proposed unsupervised methods for parallel sentence extraction that use bilingual word embeddings induced in an unsupervised manner. Our work is the first to explore using contextual representations (mBERT; Devlin et al., 2019) in an unsupervised manner to mine for bitext, and to

show improvements over the latest UNMT systems (Lample and Conneau, 2019; Song et al., 2019), for which transformers and encoder/decoder pre-training have doubled or tripled BLEU scores on unsupervised WMT'16 En-De since Artetxe et al. (2018) and Lample et al. (2018c).

5.2 Self-training Techniques

Self-training refers to techniques that use the outputs of a model to provide labels for its own training. Yarowsky (1995) proposed a semi-supervised strategy where a model is first trained on a small set of labeled data and then used to assign pseudo-labels to unlabeled data. Semi-supervised self-training has been used to improve sentence encoders that project sentences into a common semantic space. For example, Clark et al. (2018) proposed cross-view training (CVT) with labeled and unlabeled data to achieve state-of-the-art results on a set of sequence tagging, MT, and dependency parsing tasks.

Semi-supervised methods require some annotated data, even if it is not directly related to the target task. Our work is the first to apply *unsupervised* self-training for generating cross-lingual sentence embeddings. The most similar approach to ours is the prevailing scheme for unsupervised NMT (Lample et al., 2018c), which relies on multiple iterations of backtranslation (Sennrich et al., 2016) to create a sequence of pseudo-parallel sentence pairs with which to bootstrap an MT model.

6 Conclusion

In this work, we describe a novel approach for state-of-the-art unsupervised bitext mining using multilingual contextual representations. We extract pseudo-parallel sentences from unaligned corpora to create models that achieve state-of-the-art performance on unsupervised and low-resource translation tasks. Our approach is complementary to the improvements derived from initializing MT models with pre-trained encoders and decoders, and helps narrow the gap between unsupervised and supervised MT. We focused on mBERT-based embeddings in our experiments, but we expect unsupervised self-training to improve the unsupervised bitext mining and downstream UNMT performance of other forms of multilingual contextual embeddings as well.

Our findings are in line with recent work showing that multilingual embeddings are very useful for cross-lingual zero-shot and zero-resource tasks. Even without using aligned corpora, mBERT can embed sentences across different languages in a consistent fashion according to their semantic content. More work will be needed to understand how contextual embeddings discover these cross-lingual correspondences.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful comments.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. **DOI:** https://doi.org/10.18653/v1/D18-1399

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203. Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610. **DOI:** https://doi.org/10.1162/tacl_a_00288

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora.

In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604. Ann Arbor, Michigan. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1219840.1219914

Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 2–14. Da Nang, Vietnam.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925. Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194. Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-1224

Francis Grégoire and Philippe Langlais. 2018. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1442–1453. Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jian Guo, He He, Tong He, Leonard Lausen, Mu Li, Haibin Lin, Xingjian Shi, Chenguang Wang, Junyuan Xie, Sheng Zha, Aston Zhang, Hang Zhang, Zhi Zhang, Zhongyue Zhang, Shuai Zheng, and Yi Zhu. 2020. GluonCV and GluonNLP: Deep learning in computer vision and natural language processing. *Journal of Machine Learning Research*, 21:23:1–23:7.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176. Brussels, Belgium. Association for Computational Linguistics.

Viktor Hangya, Fabienne Braune, Yuliya Kalasouskaya, and Alexander Fraser. 2018. Unsupervised parallel sentence extraction from comparable corpora. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 7–13. Bruges, Belgium.

Viktor Hangya and Alexander Fraser. 2019. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234. Florence, Italy. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P19-1118

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The Sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207. Boston, MA. Association for Machine Translation in the Americas.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Melbourne, Australia. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *CoRR*, abs/1702.08734v1. **DOI:** https://doi.org/10.1109/TBDATA.2019.2921572

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. 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 1355–1360. Hong Kong, China. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252. Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W18-6325

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739. Belgium, Brussels. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W18-6453

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

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, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. 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. 2018c. 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.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual BERT? *CoRR*, abs/1911.03310v1.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92. Brussels, Belgium. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 76–79. Da Nang, Vietnam.

Benjamin Marie, Hour Kaing, Aye Myat Mon, Chenchen Ding, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. Supervised and unsupervised machine translation for Myanmar-English and Khmer-English. In *Proceedings of the 6th Workshop on Asian Translation*, pages 68–75. Hong Kong, China. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D19-5206

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations

improve crosslingual transfer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3912–3918. Minneapolis, Minnesota. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N19-1392

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Workshop on Spoken Language Translation*. Hong Kong, China. Zenodo.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W18-6319

Philip Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. David Farwell, Laurie Gerber, and Eduard H. Hovy, editors, In *Machine Translation and the Information Soup, Third Conference of the Association for Machine Translation in the Americas, AMTA '98, Langhorne, PA, USA, October 28-31, 1998, Proceedings*, volume 1529 of *Lecture Notes in Computer Science*, pages 72–82. Springer.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380. **DOI:** https://doi.org/10.1162/089120103322711578

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234. Melbourne, Australia. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P18-2037

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. *CoRR*, abs/1907.05791v2.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. CCMatrix: Mining billions of high-quality parallel sentences on the WEB. *CoRR*, abs/1911.04944v2.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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. **DOI:** https://doi.org/10.18653/v1/P16-1009

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. 2019. Understanding and improving layer normalization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 4383–4393.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55. Melbourne, Australia. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/P18-1005

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196. Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Jiawei Zhou and Phillip Keung. 2020. Improving non-autoregressive neural machine translation with monolingual data. In *ACL*.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67. Vancouver, Canada. Association for Computational Linguistics.