Boosting Unsupervised Machine Translation with Pseudo-Parallel Data

Ivana Kvapilíková kvapilikova@ufal.mff.cuni.cz
Ondřej Bojar bojar@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, 118 00, Czechia

Abstract

Even with the latest developments in deep learning and large-scale language modeling, the task of machine translation (MT) of low-resource languages remains a challenge. Neural MT systems can be trained in an unsupervised way without any translation resources but the quality lags behind, especially in truly low-resource conditions. We propose a training strategy that relies on pseudo-parallel sentence pairs mined from monolingual corpora in addition to synthetic sentence pairs back-translated from monolingual corpora. We experiment with different training schedules and reach an improvement of up to 14.5 BLEU points (English to Ukrainian) over a baseline trained on back-translated data only.

1 Introduction

After the great advancements in machine translation (MT) quality brought by neural MT (NMT; Bahdanau et al., 2015; Vaswani et al., 2017) trained on millions of pre-translated sentence pairs, there came a realization that parallel data is expensive and surely not available for most language pairs in the world. Researchers started focusing their attention on methods leveraging monolingual data for machine translation (Sennrich et al., 2016b) and even explored the extreme scenario of training a translation system in a completely unsupervised way with no parallel data at all (Artetxe et al., 2018b; Lample et al., 2018a).

The recent impressive progress in language modeling did not leave the area of machine translation intact. However, the translation capabilities of large language models such as the latest GPT models (Brown et al., 2020) are weak for underrepresented languages (Hendy et al., 2023) and unsupervised MT aimed at low-resource languages still deserves special attention.

There are two ways to approach machine translation trained exclusively on monolingual data. In the absence of parallel texts, the monolingual training sentences can either be coupled with their synthetic counterparts which are automatically generated through back-translation (Artetxe et al., 2018b; Lample et al., 2018a), or with authentic counterparts which are automatically selected from existing monolingual texts to be as close translations as possible (Ruiter et al., 2019). Researchers have successfully explored both of these avenues with the conclusion that it is indeed possible to train a functional MT system on monolingual texts only. However, little attention has been paid to combining the two approaches together.

In this paper, we work with the standard framework for training unsupervised MT but we incorporate an additional training step where sentence pairs mined from monolingual corpora are used to train the model with a standard supervised MT objective. We consider the mined

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 135–147
September 4–8, 2023, Macau SAR, China.

sentence pairs as *pseudo-parallel* as they should ideally be identical in meaning but in practice only share a certain degree of similarity. We show that they improve the translation quality nonetheless. We experiment with different training schedules to determine when to incorporate the pseudo-parallel data and when to remove it from the training.

In Section 2, we summarize the related work on the topics of unsupervised MT and parallel corpus mining. In Section 3, we introduce our method, focusing on how we obtain the pseudo-parallel sentences and how we incorporate them into the unsupervised MT training. Section 4 gives the results of our experiments which are discussed in Section 5.

2 Related Work

We separate two lines of work in the area of low-resource MT: unsupervised training on monolingual data where the research focuses on the training techniques (*unsupervised MT*) and supervised training on mined parallel sentences where the research focuses on how to create the training corpus (*parallel corpus mining*).

2.1 Unsupervised MT

Unsupervised MT was first tackled by Artetxe et al. (2018b) and Lample et al. (2018a) who introduced a neural model with shared encoder parameters for both language directions that was capable of translating without being trained on parallel data. The authors relied on pre-trained embeddings to ignite the learning process and then trained the model using denoising (Vincent et al., 2008) and back-translation (Sennrich et al., 2016a). Artetxe et al. (2018a) and Lample et al. (2018a) also explored the possibilities of unsupervised phrase-based MT where the initial phrase table is induced from a cross-lingual embedding space.

A significant improvement in neural models was brought by splitting the training of the entire model into a pre-training phase where the weights are first trained on an auxiliary task aimed at language understanding (e.g. masked language modeling, denoising) and a fine-tuning phase where the model is trained for translation. Conneau and Lample (2019) train a cross-lingual BERT-like (Devlin et al., 2018) language model on the concatenation of the monolingual corpora and copy its weights to initialize the parameters of both the encoder and the decoder. Song et al. (2019) reach slightly better translation quality by pre-training the entire sequence-to-sequence model to reconstruct a missing piece of a sentence given the surrounding tokens.

Liu et al. (2020) explore the benefits of multilingual pre-training of the entire translation model on the task of multilingual denoising (mBART) and reach state-of-the-art results in unsupervised MT. Üstün et al. (2021) extend the pre-trained mBART model with denoising adapters and fine-tune on auxiliary parallel language pairs without the need for back-translation. Garcia et al. (2020, 2021) train a multilingual translation system and combine back-translation from monolingual data with cross-translation of auxiliary parallel data in high-resource language pairs.

Unsupervised MT has been influenced by the latest advancements in large-scale multilingual language modeling (Costa-jussà et al., 2022). The GPT-3 model (Brown et al., 2020) is capable of translation without being trained on an explicit translation objective and its performance increases considerably with one-shot or few-shot fine-tuning. However, its ability to handle low-resource and non-English-centric language pairs lags behind (Hendy et al., 2023).

2.2 Parallel Corpus Mining for MT

Using mined sentence pairs for MT training was heavily explored by Schwenk (2018) and Artetxe and Schwenk (2019b) who introduced LASER, a multilingual sentence encoder that is able to find translation equivalents in 93 languages with high precision. Costa-jussà et al. (2022) extend the approach to cover 200 languages by student-teacher training. However, the training of the teacher model is heavily supervised by millions of parallel sentence pairs and its distillation also requires at least some parallel sentences.

Ruiter et al. (2019) introduce self-supervised translation where the model used for selecting translation examples is the emergent NMT model itself. The authors search for the nearest neighbors in a sentence embedding space extracted from an NMT system and apply a strong filter to only select meaningful candidates for training. Tran et al. (2020) use self-supervised training of a pre-trained multilingual model (mBART) which iteratively selects parallel sentence pairs and trains itself on the mined examples. They show an improvement over the mBART model fine-tuned on back-translated data only.

Similar to our work, Ruiter et al. (2021) incorporate a training step using denoising and back-translation into their self-supervised MT system. We take the opposite direction to reach a similar goal when we start from an unsupervised MT system and incorporate a training step supervised by the mined sentence pairs extracted outside of the NMT model. Kvapilíková and Bojar (2022) observed a positive role of pseudo-parallel data in an unsupervised MT shared task but the most effective way to integrate this type of data into the training is yet to be established.

3 Unsupervised MT with Pseudo-Parallel Data

It was demonstrated by Artetxe et al. (2018b) and Lample et al. (2018a) that the key elements of an unsupervised neural MT are shared model parameters, good initialization, and iterative learning on back-translated data. We build upon the existing work in unsupervised MT and extend the training procedure with a training step leveraging pseudo-parallel sentence pairs obtained from monolingual training corpora.

3.1 Search for Pseudo-Parallel Data

A multilingual language model trained on monolingual data only can be used to create language-neutral sentence representations (Libovický et al., 2020) in an unsupervised way. Pseudo-parallel sentence pairs are retrieved as closest neighbors in the multilingual space (Artetxe and Schwenk, 2019a).

Sentence Encoder

Multilingual masked language models (MLMs) such as mBERT (Devlin et al., 2018), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2019) are Transformer (Vaswani et al., 2017) encoders trained with a masked language modeling (MLM) objective (Devlin et al., 2018) where random tokens from the input text stream are masked and the model is trained to predict them back. MLM models create representations where each token carries information about its left and right context. Sentence embeddings can be retrieved from any layer of the model but the per-token encoder outputs need to first be aggregated, e.g. by taking their mean or their element-wise maximum over the sentence tokens.

Pires et al. (2019) and Libovický et al. (2020) studied the language neutrality of the representations produced by multilingual language models and Kvapilíková et al. (2020) showed that with minimal fine-tuning, the sentence embeddings extracted from the mid-layers of the model by mean-pooling per-token encoder outputs can be used for parallel corpus mining. They also observed that fine-tuning an MLM sentence encoder on a small synthetic parallel corpus increases both precision and recall on the task of parallel sentence mining even for unrelated language pairs.

Parallel Sentence Search

To perform the search for parallel sentence pairs, all sentences from the two monolingual corpora are encoded and all possible sentence combinations are scored to select the most similar sentence pairs. The scoring is performed by a margin-based similarity metric (Artetxe and Schwenk, 2019a)

$$\mathrm{xsim}(x, y) = \mathrm{margin}\Big(\cos(x, y), \sum_{z \in \mathrm{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \mathrm{NN}_k(y)} \frac{\cos(y, z)}{2k}\Big) \tag{1}$$

where $\mathrm{margin}(a, b) = \frac{a}{b}$, $\mathrm{NN}_k(x)$ is the set of k nearest neighbors of x. The method for scoring involves cosine similarity which is comparatively evaluated against the average cosine similarity of a given sentence with its nearest neighbors to eliminate the "hubs". When the score surpasses a designated threshold T, two sentences are deemed to be parallel:

$$\mathrm{xsim}(x, y) > T \tag{2}$$

3.2 Unsupervised MT Architecture

The design of an NMT system needs to meet several requirements to be functional for unsupervised translation. Firstly, a significant number of parameters needs to be shared among the languages in order to allow the model to generate a shared latent space where meaning is represented regardless of the language it is expressed in (Lample et al., 2018b). Secondly, the initialization of the model weights is vital to produce an initial solution and kick-start the training process (Conneau and Lample, 2019).

The configuration of our unsupervised MT system follows that of Conneau and Lample (2019) and consists of a Transformer encoder and decoder, both of which are shared between the two languages. The tokenized input in both languages is processed by a single BPE (Sennrich et al., 2016b) model learned on the concatenation of the two monolingual corpora and the joint vocabulary enables both languages to use a shared embedding matrix.

3.3 Unsupervised Pre-Training

The model is initialized with weights from a masked language model pre-trained on the monolingual corpora and copied into both the encoder and the decoder as in Conneau and Lample (2019). The initialized model is further pre-trained as a bilingual denoising autoencoder (Liu et al., 2020). The fine-tuning of the pre-trained model is scheduled in stages which are discussed in Section 3.4.

3.4 Fine-Tuning for Translation

The pre-trained model is fine-tuned on both back-translated and pseudo-parallel data which are combined into different training schedules to determine their role at a given point in training. Intuitively, non-equivalent sentence pairs with some translation information should be useful at the beginning of the training when the model has minimal or no cross-lingual information. However, as the training progresses, it starts to produce synthetic translations of increasing quality which at a certain point surpass the quality of the pseudo-parallel corpus. We hypothesize that the most effective approach is to train the model on both synthetic and pseudo-parallel data until a certain breaking point, and from that point on, continue training solely on synthetic data.

3.4.1 Fine-Tuning on Pseudo-Parallel Data

To fine-tune the model on pseudo-parallel data, the standard supervised MT objective is used. In every step of the training, a mini-batch of pseudo-parallel sentences is added and the model is trained to minimize the loss function

$$L_{PPMT}(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}) = E_{(x,y) \sim \textit{PseudoPar}, \hat{y} \sim \mathrm{dec}(\mathrm{enc}(x))} \Delta(\hat{y}, y) \tag{3}$$

	de-hsb	**en-ka**	**en-kk**	**en-uk**
train (mono)	29.4M/0.9M	17.1M/6.6M	17.1M/7.7M	17.1M/17.3M
train (pseudo-parallel)	770K	230K	169K	496K

Table 1: Number of sentences in the monolingual corpora and mined pseudo-parallel corpora.

where $(\theta_{\text{enc}}, \theta_{\text{dec}})$ is the trained model, (x, y) is a sentence pair sampled from the pseudo-parallel data set $PseudoPar$, and Δ is the cross-entropy loss.

3.4.2 Fine-Tuning on Iteratively Back-Translated Data

In the back-translation step, the model is first set to the inference mode and used to translate a batch of sentences. The synthetic translations serve as source sentences fed into the model while the original sentences serve as the ground truth for the cross-entropy loss computation. The back-translation loss for translation from language $Lsrc$ to $Ltgt$ is defined as

$$L_{IBT}(\theta_{\text{enc}}, \theta_{\text{dec}}, Ltgt) = E_{x \sim D_{Ltgt}, \hat{x} \sim \text{dec}(\text{enc}(T(x)))}(\Delta(\hat{x}, x)) \qquad (4)$$

where x is a sentence sampled from the target corpus D_{Ltgt}, $T(x)$ is the translation model which generates a synthetic translation of x, and Δ is the cross-entropy loss.

4 Experimental Details

4.1 Data

We train translation models for the following language pairs: German-Upper Sorbian (de-hsb), English-Georgian (en-ka), English-Kazakh (en-kk) and English-Ukrainian (en-uk). The German and Upper Sorbian monolingual training data as well as the parallel validation and test sets were provided in the WMT22 unsupervised shared task (Weller-Di Marco and Fraser, 2022). The monolingual training data for the other languages come from the Oscar[1] corpus. The training data summary is given in Table 1. The English-centric validation and test sets were taken from the Flores Evaluation Benchmark (Costa-jussà et al., 2022). In addition, the legal test sets from the MT4All shared task (de Gibert Bonet et al., 2022) were used for evaluation.

The data was tokenized and split into BPE units using the fastText (Joulin et al., 2016) library. We shared one BPE vocabulary of 55k entries for en-ka-kk-uk and another vocabulary of 18k entries for de-hsb.

4.2 Training Details

4.2.1 Model Architecture

All our translation models have a dual character to translate in both translation directions. They have the same 6-layer Transformer architecture with 8 attention heads and the hidden size of 1024, language embeddings, GELU (Hendrycks and Gimpel, 2017) activations and a dropout rate of 0.1. For language model pre-training, we use mini-batches of 64 text streams (256 tokens per stream) per GPU and Adam (Kingma and Ba, 2015) optimization with `lr=0.0001`. For denoising and MT fine-tuning, we use mini-batches of 3400 tokens per GPU and Adam optimization with a linear warm-up `(beta1=0.9,beta2=0.98,lr=0.0001)`. The models are trained on 8 GPUs. We use the XLM[2] toolkit for training.

4.2.2 Sentence Encoder

We use the XLM-100 model (Conneau and Lample, 2019) fine-tuned on English-German synthetic sentence pairs according to Kvapilíková et al. (2020) as our sentence encoder. To mea-

[1] https://oscar-project.org/

[2] https://github.com/facebookresearch/XLM

	de-hsb	**en-ka**	**en-kk**	**en-uk**
Precision	87.08	44.8	49.3	67.4
Recall	76.15	44.4	42.4	74.2
F1	81.25	44.6	45.6	70.6
Threshold	1.034	1.023	1.022	1.026

Table 2: The evaluation metrics on the PSM task and the respective mining thresholds.

sure its ability to create representations with a high level of multilingualism, we evaluate its performance of an auxiliary task of parallel sentence mining (PSM). For each language pair, we randomly select 200k sentences from the monolingual data, mix in the parallel validation set, and measure the precision and recall of the model when trying to reconstruct it.

Since XLM-100 was trained on 100 languages and Upper Sorbian is not one of them, we fine-tune the model on German and Upper Sorbian sentences before using it to mine parallel sentence pairs. We stop fine-tuning when the quality of the mined corpus starts deteriorating. We determine the optimal length of fine-tuning on the PSM task and observe that both precision and recall start slowly decreasing after the model had seen 500k sentences.

To retrieve sentence embeddings from the trained model, we mean-pool the encoder outputs from the fifth-to-last layer across sentence tokens (the layer and aggregation choice follow Kvapilíková et al. (2020)). We search the embedding space as described in Equation (1) and Equation (2). We select a threshold T that maximizes the F1 score on the PSM task. Table 2 lists the precision and recall of all sentence encoders used for mining together with the optimal mining threshold. The amount of mined parallel sentences used for unsupervised MT training is given in Table 1.

4.2.3 Pre-Training

We pre-train one multilingual language model for en+ka+kk+uk and one bilingual language model for de+hsb. In one training step, the model sees a minibatch of text streams in all languages. The weights from the pre-trained language models are copied into both the encoder and the decoder of the respective bilingual NMT models. The initialized NMT model for each language pair is then further pre-trained with the denoising auto-encoding loss on the two languages until convergence. The details of the denoising task are identical to Lample et al. (2018a).

4.2.4 Fine-Tuning

We experiment with different fine-tuning strategies for unsupervised machine translation. For each language pair, all translation models are initialized with the same weights obtained in the pre-training stage described in the previous paragraph.

IBT (baseline) models are fine-tuned solely with the iterative back-translation loss.

PseudoPar models are fine-tuned with the standard supervised MT loss on our pseudo-parallel corpora.

IBT+PseudoPar models are fine-tuned simultaneously with the iterative back-translation loss on the monolingual sentences and with the standard MT loss on the pseudo-parallel sentence pairs.

IBT+PseudoPar$\mapsto$IBT models are a continuation from different checkpoints of the *IBT+PseudoPar* models where the supervised MT objective is dropped and the training continues with iterative back-translation only. We experiment with different checkpoints to find the optimal point to switch the training.

	de-hsb	**hsb-de**	**en-ka**	**ka-en**	**en-kk**	**kk-en**	**en-uk**	**uk-en**
WMT22 best	17.9	18.0	-	-	-	-	-	-
ChatGPT	6.4	-	3.9	-	5.2	-	**25.8**	-
IBT (baseline)	29.5	35.6	3.6	5.2	0.8	1.0	8.4	12.9
PseudoPar	11.3	12.0	1.9	4.8	1.0	3.1	4.6	8.6
IBT+PseudoPar	32.18	36.13	6.8	12.7	5.9	11.3	12.2	20.8
↦IBT	**34.94**	**39.63**	**7.7**	**14.0**	**7.2**	**12.1**	**15.7**	**23.7**

	de-hsb	**hsb-de**	**en-ka**	**ka-en**	**en-kk**	**kk-en**	**en-uk**	**uk-en**
de Gibert Bonet (2022)	-	-	12.0	-	6.4	-	20.8	-
IBT (baseline)	-	-	9.0	12.7	0.3	0.3	14.9	12.6
PseudoPar	-	-	2.1	6.8	8.0	11.6	14.6	13.1
IBT+PseudoPar	-	-	11.5	22.0	**16.3**	**18.6**	**29.3**	21.7
↦IBT	-	-	**15.0**	**23.5**	9.3	12.7	27.5	**21.8**

Table 3: MT performance of our systems measured by BLEU scores on the general test set (top) and the legal test set (bottom). Compared to the WMT22 winner (Shapiro et al., 2022), ChatGPT, and the system trained by de Gibert Bonet et al. (2022).

4.2.5 Evaluation

The baseline for our approach is an improved model of Conneau and Lample (2019) with an extra pre-training step on the denoising task for better performance. We initialize the baseline model with the weights of a cross-lingual language model, further pre-train as a denoising autoencoder and fine-tune with iterative back-translation.

We benchmark our results against MT systems of de Gibert Bonet et al. (2022) trained as a baseline for the MT4All shared task according to the methodology of Artetxe et al. (2019), and against Shapiro et al. (2022) who won the WMT22 de-hsb unsupervised task with a multilingual system that was pre-trained according to the mBART (Liu et al., 2020) methodology and fine-tuned on synthetic texts generated by a phrase-based system.

To challenge the relevance of unsupervised MT in the world of large language models, we also translate our test sets by the GPT-3.5 Turbo model[3] using the ChatGPT API and compare to our results.

We measure translation quality by BLEU score using sacreBLEU[4] (Post, 2018).

5 Results & Discussion

5.1 Results

We observed a significant improvement in translation quality over the baseline for all translation pairs. Table 3 shows that the baseline *IBT* system falls short of our proposed method by between 4.7 BLEU points (en→kk) and 10.7 BLEU points (uk→en) on the general test set. The differences on the legal test set are even more pronounced: we observe an increase of up to 14.5 BLEU over the baseline (en→uk). Our de→hsb system outperforms the WMT22 winner by 17 BLEU points. When translating from English to Kazakh, our approach reaches a BLEU score of 16.3 while the baseline which solely relies on iterative back-translation does not receive enough cross-lingual signal to start learning at all. The hybrid system by de Gibert Bonet et al. (2022) which uses additional translation information from an unsupervised phrase-based system falls behind with a BLEU score of 6.4.

[3] `https://platform.openai.com/docs/models/gpt-3-5`

[4] `sacrebleu -tok '13a' -s 'exp'`

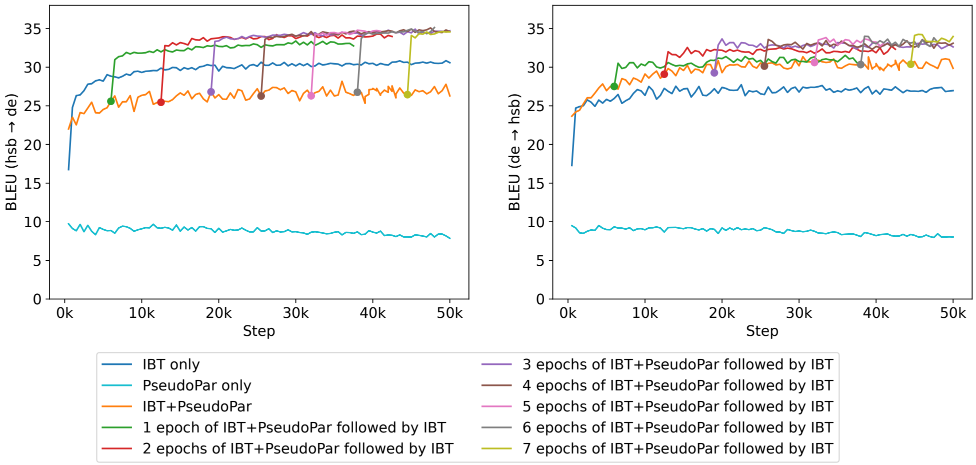

Figure 1: The development of validation BLEU scores during training. Any parallel resources were prohibited.

The results of translation by ChatGPT from English or German into truly low-resource languages (hsb, ka, kk) are significantly worse than our results. However, after manually evaluating several translations with a zero BLEU score, we believe that the automatic metric puts ChatGPT's less literal translations at a disadvantage. ChatGPT definitely favors fluency over accuracy, but it gets zero BLEU credit even in situations when it conveys the same information in different words. Nonetheless, the en→uk translation by ChatGPT is better than all unsupervised MT systems. It must be noted that the systems cannot be directly compared to ChatGPT since its training corpus is larger and might include parallel texts.

5.2 Training Schedules

Figure 1 shows training curves with validation BLEU scores of all our de↔hsb systems. We see that the *IBT+PseudoPar* system trained simultaneously on back-translated and pseudo-parallel data without any special schedule outperforms the baseline for de→hsb but not in the opposite direction. For hsb→de, the baseline performance is surpassed as soon as we remove the pseudo-parallel corpus from the training.

We trained several de-hsb models starting from *IBT+PseudoPar* after each completed epoch of 770k pseudo-parallel sentences. Upon examination of the training curves in Figure 1, we see an immediate increase in validation BLEU score of ∼0.9–4.9 BLEU points which occurred within the first 500 training steps after removing the pseudo-parallel corpus from the training. This observation confirms our hypothesis that pseudo-parallel sentence pairs aid the training in the beginning but the quality of the corpus itself poses an upper bound on the performance of the system. However, removing the corpus too early (after one or two epochs) leads to a lower final BLEU score. Therefore, we recommend to keep training the *IBT+PseudoPar* model until convergence and only then switch to iterative back-translation alone *IBT+PseudoPar*↦*IBT*.

The flat *PseudoPar* training curves indicate that the quality of the pseudo-parallel corpus alone is inadequate for training a functional MT system without back-translation.

5.3 Domain-specific MT

Interestingly, removing the pseudo-parallel corpus from the training harms the translation quality measured on the legal test sets where the best performance for en→kk, kk→en and en→uk

#	Upper Sorbian	German	Score
1	Thomas de Maizière	Thomas de Maizière	1.286
2	Es ist ein harter Kampf, die Konkurrenz ist groß.	To bě napjata hra, a konkurenca bě wulka.	1.185
3	Der Roman hat *1200* Seiten.	Kniha ma *300* stronow.	1.178
4	Er passt zu diesem Team wie der Deckel auf den Topf.	Wón so k mustwu hodźi kaž wěko na hornc.	1.161
5	Die größte misst über *fünf Meter, die kleinste wenige Milli*meter.	Najkrótša měri *10 cm*, najdlěša *1 meter.*	1.101
6	Wer Wohlstand will, braucht Wissenschaft.	Štóž chce *něšto změnić, trjeba sylnu wolu.*	1.063
7	*Auch für Apple ist das iPhone wichtig.*	*Tež aleje su jara važne.*	1.037

Table 4: A sample from the de-hsb mined parallel corpus. Non-matching words in italics.

is achieved by *IBT+PseudoPar*. We suspect that this is the result of the repeating terminology in the domain-specific test sets which is better handled by the *IBT+PseudoPar* for some language pairs. This is consistent with the fact that the *PseudoPar* system trained exclusively on pseudo-parallel data performs quite well on the en-kk and en-uk legal test set (8.0 on en→kk, 11.6 on kk→en and 14.6 on en→uk) while having poor results on the general test set (1.0 on en→kk, 3.1 on kk→en and 4.6 on en→uk). Based on our findings, we believe that utilizing pseudo-parallel sentences extracted from domain-specific monolingual corpora has the potential to enhance the training of domain-specific MT in general. However, further experiments are out of the scope of this paper.

5.4 Data quality

The sentence pairs in the pseudo-parallel corpus are far from equivalent in meaning. As illustrated in Table 4, many of the sentences are paired because they share a named entity, a numeral (not necessarily identical), a punctuation mark, or one distinctive word. Others have a similar sentence structure, they contain a similar segment or they contain words that are somehow related, e.g. Apple/alleys (*"aleje"*), although the word Apple is not the fruit in this context. On the other hand, synthetic sentences in the first training iterations are also extremely noisy, and even later they contain artifacts such as non-translated words or mistranslated named entities.

Table 5 shows what the back-translated and pseudo-parallel data can look like. We observed how the back-translated version of one sentence changes as the training progresses and witnessed several types of error, e.g. the German word *"laufend"* is not translated at all in the initial iterations; the word "April" remains mistranslated as "March" ("*měrc*") throughout the entire training. On the other hand, the pseudo-parallel sentence matched based on its distance from the source sentence has a similar meaning but is factually inaccurate.

We see that many of the pseudo-parallel translations are far from equivalent but it is difficult to measure the quality of the entire corpus. We measure it indirectly by the increase in BLEU score associated with introducing the corpus into the unsupervised MT training or by measuring the quality of the sentence encoder used for creating the corpus. To be able to evaluate the precision/recall of the sentence encoder, we have to control the number of parallel sentences hidden in the input corpora. However, in real-life scenarios, the level of comparability of two monolingual corpora is never known precisely. If the monolingual corpora provided for unsupervised translation come from a different domain and contain dissimilar sentences, the model has no good candidates to find. This poses a challenge especially when setting the correct mining threshold for the monolingual corpora at hand.

It is not clear what are the attributes of the pseudo-parallel corpus that the unsupervised

SRC	Ich musste mich laufend weiterbilden, und so legte ich im April 1952 die erste und ein Jahr darauf die zweite Lehramtsprüfung ab.
REF	Dyrbjach so běžnje dale kwalifikować, a tak złožich w aprylu 1952 prěnje a lěto po tym druhe wučerske pruwowanje.
PseudoPar	*Hańža Winarjec-Orsesowa* wotpołoži prěnje wučerske pruwowanje *w lěće 1949 a druhe w lěće* 1952.
IBT @ 500	Dyrbjach so *laufend* dale *kubłać*, a tak *legte w měrcu* 1952 *prěnje a lěto na to druhe Lejnjanske pruwowanje ab.*
IBT @ 3000	Dyrbjach so běžnje dale *kubłać*, a tak w *měrcu* 1952 prěnju a lěto na to druhu *lektoratu serbšćiny wotpołožichmy.*
IBT @ 10000	Dyrbjach so běžnje dale *kubłać*, a tak wotpołožich w *měrcu* 1952 prěnju a lěto *na* to druhu *lektoratu.*

Table 5: A sample sentence translated by the IBT model after 500, 3,000 and 10,000 training steps compared to the closest neighbor of such sentence from the bilingual sentence space (PseudoPar). The mistranslated words are indicated in italics.

MT training benefits from the most. We believe that the benefits of training on such noisy data are twofold: 1) the perfect matches are a valuable source of correct supervision, and 2) the abundant less-than-perfect matches still introduce a new translation signal which can help the model leave a suboptimal situation which we often observe during back-translation when the model learns to mistranslate a word and never forgets it.

6 Conclusion

We have demonstrated the benefits of MT training on pseudo-parallel data in situations when true parallel data is not available. While the pseudo-parallel corpus alone does not reach sufficient quality for standard supervised MT training, it works well in combination with iterative back-translation. It is optimal to train the model until convergence on both pseudo-parallel and synthetic sentence pairs, remove the pseudo-parallel corpus and continue training with iterative back-translation only.

Incorporating similar sentence pairs into the standard unsupervised MT training increases translation quality across all evaluated language pairs with an improvement of up to 14.5 BLEU over the baseline trained without pseudo-parallel data and 8.5 BLEU over a hybrid unsupervised system (en→uk). Furthermore, we observed that in some situations (en↔kk), the iterative back-translation becomes trapped in a suboptimal state where no learning occurs. Introducing pseudo-parallel data can rescue the model from this state and trigger the learning process.

After evaluating our approach on a legal test set, we believe that training on pseudo-parallel sentences could be particularly useful for domain-specific unsupervised MT. If we have two in-domain monolingual corpora at hand, parallel corpus mining is an efficient strategy to retrieve translation information.

The pseudo-parallel corpus helps the training despite being noisy. We hypothesize that while exact translations help the model find correct correspondences, also the noise can introduce new information and prevent the model from memorizing some of the artifacts of back-translated sentences. We leave it up to future research to evaluate whether a cleaner but smaller corpus would bring even larger gains.

Acknowledgements

This research was partially supported by the grant 19-26934X (NEUREM3) of the Czech Science Foundation and by the SVV project number 260 698 of the Charles University.

References

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on EMNLP*, Brussels. 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. Association for Computational Linguistics.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018b). Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Artetxe, M. and Schwenk, H. (2019b). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bahdanau, D., Cho, K., and Bengio, Y. (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*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.

de Gibert Bonet, O., Goenaga, I., Armengol-Estapé, J., Perez-de Viñaspre, O., Parra Escartín, C., Sanchez, M., Pinnis, M., Labaka, G., and Melero, M. (2022). Unsupervised machine translation in real-world scenarios. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3038–3047, Marseille, France. European Language Resources Association.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv [e-Print archive]*, abs/1810.04805.

Garcia, X., Foret, P., Sellam, T., and Parikh, A. P. (2020). A multilingual view of unsupervised machine translation.

Garcia, X., Siddhant, A., Firat, O., and Parikh, A. (2021). Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137, Online. Association for Computational Linguistics.

Hendrycks, D. and Gimpel, K. (2017). Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Kvapilíková, I., Artetxe, M., Labaka, G., Agirre, E., and Bojar, O. (2020). Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Kvapilíková, I. and Bojar, O. (2022). CUNI submission to MT4All shared task. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 78–82, Marseille, France. European Language Resources Association.

Lample, G., Denoyer, L., and Ranzato, M. (2018a). Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018b). Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on EMNLP*, pages 5039–5049.

Libovický, J., Rosa, R., and Fraser, A. (2020). On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, 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.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Ruiter, D., España-Bonet, C., and van Genabith, J. (2019). Self-supervised neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1828–1834, Florence, Italy. Association for Computational Linguistics.

Ruiter, D., Klakow, D., van Genabith, J., and España-Bonet, C. (2021). Integrating unsupervised data generation into self-supervised neural machine translation for low-resource languages. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 76–91, Virtual. Association for Machine Translation in the Americas.

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). 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.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin. Association for Computational Linguistics.

Shapiro, A., Salama, M., Abdelhakim, O., Fayed, M., Khalafallah, A., and Adly, N. (2022). The AIC system for the WMT 2022 unsupervised MT and very low resource supervised MT task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1117–1121, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. (2019). MASS: masked sequence to sequence pre-training for language generation. *CoRR*, abs/1905.02450.

Tran, C., Tang, Y., Li, X., and Gu, J. (2020). Cross-lingual retrieval for iterative self-supervised training. *CoRR*, abs/2006.09526.

Üstün, A., Berard, A., Besacier, L., and Gallé, M. (2021). Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. pages 1096–1103.

Weller-Di Marco, M. and Fraser, A. (2022). Findings of the wmt 2022 shared tasks in unsupervised mt and very low resource supervised mt. In *Proceedings of the Seventh Conference on Machine Translation*, pages 801–805, Abu Dhabi. Association for Computational Linguistics.