Extract and Edit: An Alternative to Back-Translation for Unsupervised Neural Machine Translation

Jiawei Wu and **Xin Wang** and **William Yang Wang**
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106 USA
{jiawei_wu,xwang,william}@cs.ucsb.edu

Abstract

The overreliance on large parallel corpora significantly limits the applicability of machine translation systems to the majority of language pairs. Back-translation has been dominantly used in previous approaches for unsupervised neural machine translation, where pseudo sentence pairs are generated to train the models with a reconstruction loss. However, the pseudo sentences are usually of low quality as translation errors accumulate during training. To avoid this fundamental issue, we propose an alternative but more effective approach, **extract-edit**, to extract and then edit real sentences from the target monolingual corpora. Furthermore, we introduce a comparative translation loss to evaluate the translated target sentences and thus train the unsupervised translation systems. Experiments show that the proposed approach consistently outperforms the previous state-of-the-art unsupervised machine translation systems across two benchmarks (English-French and English-German) and two low-resource language pairs (English-Romanian and English-Russian) by more than 2 (up to 3.63) BLEU points.

1 Introduction

Promising results have been achieved in Neural Machine Translation (NMT) by representation learning (Cho et al., 2014b; Sutskever et al., 2014). But recent studies (Koehn and Knowles, 2017; Isabelle et al., 2017; Sennrich, 2017) highlight the overreliance of current NMT systems on large parallel corpora. In real-world cases, the majority of language pairs have very little parallel data, so the models need to leverage monolingual data to address this challenge (Gulcehre et al., 2015; Zhang and Zong, 2016; He et al., 2016; Yang et al., 2018).

While many studies have explored how to use the monolingual data to improve translation performance with limited supervision, latest approaches (Lample et al., 2018a; Artetxe et al., 2018; Lample et al., 2018b) focus on the fully unsupervised scenario. Back-translation has been dominantly used in these approaches, where pseudo sentence pairs are generated to train the translation systems with a reconstruction loss. However, it is inefficient because the generated pseudo sentence pairs are usually of low quality. During the dual learning of back-translation, the errors could easily accumulate and thus the learned target language distribution would gradually deviate from the real target distribution. This critical drawback hinders the further development of the unsupervised NMT systems.

An alternative solution is to extract real parallel sentences from comparable monolingual corpora, and then use them to train the NMT systems. Recently, neural-based methods (Chu et al., 2016; Grover and Mitra, 2017; Grégoire and Langlais, 2018) aim to select potential parallel sentences

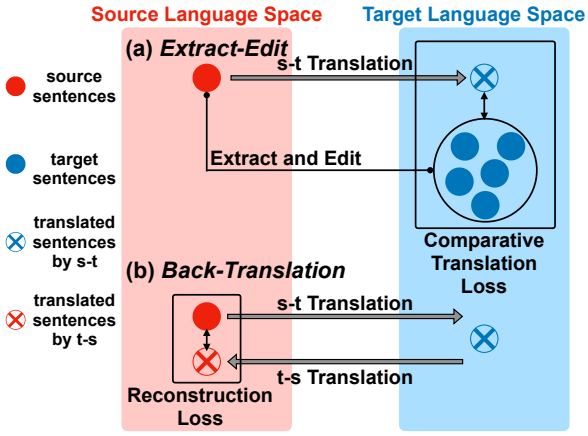

Figure 1: The comparison between two approaches of unsupervised NMT, extract-edit and back-translation. When training the source-to-target (s-t) translation model, instead of using the t-s back-translated sentences to train the model, we directly set the extracted-edited sentences as pivotal points to guide the training.

from monolingual corpora in the same domain. However, these neural models need to be trained on a large parallel dataset first, which is not applicable to language pairs with limited supervision.

In this paper, we propose a radically different approach for unsupervised NMT—**extract-edit**, a powerful alternative to back-translation (see Figure 1). Specifically, to train the source-to-target translation model, we first extract potential parallel sentence candidates in the target language space given a source language sentence. Since it cannot be guaranteed that there always exist potential parallel sentence pairs in monolingual corpora, we further propose a simple but effective editing mechanism to revise the extracted sentences, making them aligned with the source language sentence. Then a comparative translation loss is introduced to evaluate the translated sentence based on the extracted-and-edited ones and train the translation model. Compared to back-translation, extract-edit avoids the distribution deviation issue by extracting and editing real sentences from the target language space. Those extracted-and-edited sentences serve as pivotal points in the target language space to guide the unsupervised learning. Thus, the learned target language distribution could be closer to the real one. The extract-edit model and the translation model, the two major parts of our method, can be jointly trained in a fully unsupervised way.

Empirical results on popular benchmarks show that exact-edit consistently outperforms the state-of-the-art unsupervised NMT system (Lample et al., 2018b) with back-translation across four different languages pairs. In summary, our main contributions are three-fold[1]:

- We propose a more effective alternative paradigm to back-translation, extract-edit, to train the unsupervised NMT systems with potentially real sentence pairs;
- We introduce a comparative translation loss for unsupervised learning, which optimizes the translated sentence by maximizing its relative similarity with the source sentence among the extracted-and-edited pairs;
- Our method advances the previous state-of-the-art NMT systems across four different language pairs under monolingual corpora only scenario.

[1]The source code can be found in this repository: https://github.com/jiaweiw/Extract-Edit-Unsupervised-NMT

2 Background

Without parallel sentence pairs as constraints on mapping language spaces, training NMT systems is an ill-posed problem because there are many potential mapping solutions. Nevertheless, some promising methods have been proposed in this field (Lample et al., 2018a; Artetxe et al., 2018; Lample et al., 2018b). The main technical protocol of these approaches can be summarized as three steps: *Initialization*, *Language Modeling*, and *Back-Translation*. In this section, we mainly introduce the three steps and the crucial settings that we have followed in our work.

In the remainder of the paper, we denote the space of source and target languages by $\mathcal{S}$ and $\mathcal{T}$, respectively. enc and dec refer to the encoder and decoder models in the sequence-to-sequence systems. $V_{s\to t}$ stands for the composition of enc in the source language and dec in the target language, which can be viewed as the source-to-target translation system.

Initialization Given the ill-posed nature of the unsupervised NMT task, a suitable initialization method can help model the natural priors over the mapping of two language spaces we expect to reach. There are mainly two initialization methods: (1) bilingual dictionary inference (Conneau et al., 2018; Artetxe et al., 2018; Lample et al., 2018a) and (2) byte-pair encoding (BPE) (Sennrich et al., 2016b; Lample et al., 2018b). As shown in Lample et al. (2018b), the inferred bilingual dictionary can provide a rough word-by-word alignment of semantics, and the BPE can reduce the vocabulary size and eliminate the presence of unknown words in the output results.

In our *extract-edit* approach, to extract potential parallel sentence pairs, we need to compare the semantic similarity of sentences between two languages first. A proper initialization can also help align the semantic spaces and extract potential parallel pairs within them. Thus, following the previous methods, we use the inferred bilingual dictionary as described in Conneau et al. (2018) for unrelated language pairs and the shared BPE in Lample et al. (2018b) as initialization for related ones.

Language Modeling After a proper initialization, given large amounts of monolingual data, we

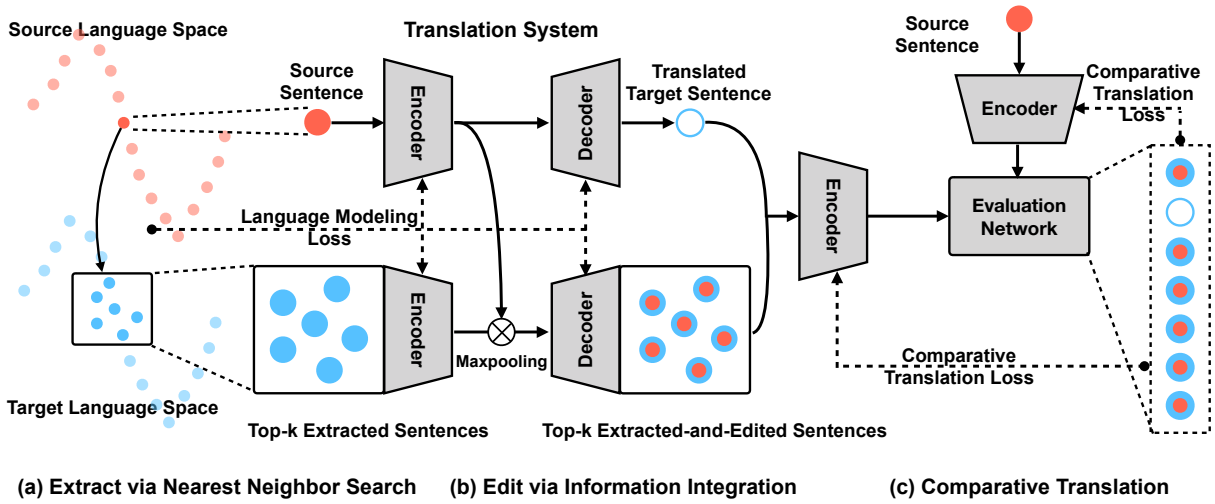

Figure 2: The overview of our unsupervised NMT model based on the *extract-edit* approach. Given a source sentence, (a) the top-k potential parallel sentences of the target language are extracted via nearest neighbor search. (b) The extracted sentences are further edited with the source sentence. (c) The evaluation network evaluates the translated sentence and the extracted-and-edited sentences based on their similarities with the source sentence. Note that (1) all the encoders share the same parameters (same for decoders); (2) the decoding processes are non-differentiable, so the language modeling loss and the comparative translation loss are used to train the learning modules before and after the decoding processes, respectively.

can train language models on both source and target languages. These models express a data-driven prior about the composition of sentences in each language. In NMT, language modeling is accomplished via denosing autoencoding, by minimizing:

$$\mathcal{L}_{lm}(\theta_{enc}, \theta_{dec}) = \mathbb{E}_{x\sim\mathcal{S}}[-\log V_{s\to s}(x|C(x))] + \mathbb{E}_{y\sim\mathcal{T}}[-\log V_{t\to t}(y|C(y))] \quad (1)$$

where C is a noise model with some words dropped and swapped, θ_{enc} and θ_{dec} are the learnable parameters of enc and dec. $V_{s\to s}$ and $V_{t\to t}$ are the encoder-decoder language models on the source and target sides, respectively.

In our *extract-edit* approach, we follow similar settings and adopt the noise model proposed by Lample et al. (2018a). Note that the parameters of all enc are shared (same for dec) in our framework to ensure a strong alignment and mapping between two languages. This sharing operation is essential for both the translation model and the extract-edit model. Thus, we use enc to represent encoders in source language modeling $enc_{\mathcal{S}}$ and in target language modeling $enc_{\mathcal{T}}$ (same for dec).

Back-Translation Back-translation (Sennrich et al., 2016b) has been dominantly used in prior work to train the unsupervised NMT system. It couples the source-to-target translation model with a backward target-to-source model and trains the whole system with a reconstruction loss. This can be viewed as converting the unsupervised problem into a supervised scenario by generating pseudo language pairs (He et al., 2016).

Despite the popularity of back-translation in the previous methods (Lample et al., 2018a; Artetxe et al., 2018; Lample et al., 2018b), we argue that it suffers from the low-quality pseudo language pairs. Thus, in this work, we propose a new paradigm, extract-edit, to address this issue by extracting and editing potential real parallel sentences. Below we describe our approach in details.

3 Extract-Edit

The overview of our *extract-edit* approach is shown in Figure 2. We first extract and edit real sentences from the target language space according to their similarities with the source sentence. These extracted-and-edited sentences serve as pivotal points in the target language space, which locate a probable region where the real target sentence could be. Then we introduce a comparative translation loss to evaluate the translated sentence and train the system. Basically, the comparative translation loss encourages the translated sentence to approximate the real sentence by maximizing its

relative similarity with the source sentence compared to the extracted-and-edited sentences. As a result, we manage to minimize the deviation of the learned target language distribution and the mapping noises between two language spaces.

3.1 Extract

Most existing methods in comparable corpora mining introduce two encoders to represent sentences of two languages separately, and then use another network to measure the similarity (Chu et al., 2016; Grover and Mitra, 2017; Grégoire and Langlais, 2018). However, owing to the shared encoders and decoders in language modeling, the semantic spaces of two languages are already strongly connected in our scenario.

Therefore, to avoid extra computation resources, we directly use the enc in language modeling to obtain sentence embeddings for two languages. As shown in Figure 2 (a), for a given source sentence s, we use the nearest neighbor search based on L_2 distance to find top-k real sentences from the target language space (k is a hyper-parameter decided empirically). The sentence embeddings used for searching are computed based on the shared encoder enc. The reason to choose top-k sentences rather than top-1 is to keep a high recall rate and obtain more related samples from the target language space. Finally, given the source sentence s, we denote M as a set of the k potential parallel target sentences:

$$M = \{t | \min_{1,\cdots,k}(||e_s - e_t||), t \in \mathcal{T}\}, \quad (2)$$

where e_s and e_t are sentence embeddings encoded by the shared encoder enc.

3.2 Edit

Even though the extracted sentences could serve as pivotal points to guide NMT, there is no guarantee that there always exists a parallel sentence in the target corpus. Thus, in order to make it closer to the real paired sentence in the target language space, we propose an editing mechanism to revise the extracted target sentence $t \in M$ based on the semantics of the source sentence s. As described in Figure 2 (b), we employ a maxpooling layer to reserve the more significant features between the source sentence embedding e_s and the extracted sentence embedding e_t ($t \in M$), and then decode it into a new sentence t':

$$M' = \{t' | t' = dec(\text{maxpooling}(e_s, e_t)), t \in M\}, \quad (3)$$

where M' is the set of the extracted-and-edited sentences. Based on the semantic information of the source sentence s, we can further improve the extracted results with this editing mechanism. Unlike other studies using the editing to generate more structural sentences (Guu et al., 2018; Hashimoto et al., 2018), here the revised sentences are designed to serve as better pivotal points in the target language space to guide the translation procedure. This can also be viewed as adding constraints when aligning the two language spaces.

3.3 Evaluate

Given a source sentence s, we can translate it as t^* using the source-to-target translation model $P_{s\to t}$. Meanwhile, a set M' of k sentences can also be generated by the extract-edit approach described above. Although the M' may contain potential parallel sentences t' for s, we cannot directly use (s, t') as ground-truth sentence pairs to train the translation model $V_{s\to t}$ because the NMT system is sensitive to noises (Cho et al., 2014a; Cheng et al., 2018). The rough operation like this will result in sub-optimal translation performance.

Therefore, in order to assess the quality of the translated sentence t^* and train the translation model $V_{s\to t}$, we introduce an evaluation network R for evaluating the relative similarities between the source and target sentences among all sentence pairs. The evaluation network R is a multilayer perceptron; it takes the target sentence embedding e_t and source sentence embedding e_s as inputs, and converts them into the joint embedding space as r_t and r_s. So the similarity

$$\alpha(t|s) = cosine(r_t, r_s) = \frac{r_t \cdot r_s}{||r_t||||r_s||}. \quad (4)$$

Then, a softmax-like formulation is used to compute the ranking score for the translated sentence t^* given the extracted-and-edited sentence set M':

$$P(t^*|s, M') = \frac{exp(\lambda\alpha(t^*|s))}{\sum_{t'\in M'\cup\{t^*\}} exp(\lambda\alpha(t'|s))}, \quad (5)$$

where the hyper-parameter λ is similar to the inverse temperature of the softmax function. Lower λ encourages the model to treat all extracted-edited sentences equally, while higher λ highlights the importance of sentences with higher-score.

3.4 Learning

Comparative Translation As introduced above, the ranking score calculates the relative similarity between the $< s, t^* >$ pair and all the extracted-and-edited pairs $< s, t' >$. Assuming we have a good evaluation network R with θ_R denoting its parameters, we further introduce the comparative translation loss $\mathcal{L}_{com}$ for unsupervised machine translation:

$$\mathcal{L}_{com}(\theta_{enc}|\theta_R) = -\mathbb{E}(\log P(t^* = V_{s\to t}(s)|s, M')), \tag{6}$$

where θ_{enc} is the parameters of the shared encoder enc. Basically, the translation model is trying to minimize the relative distance of the translated sentence t^* to the source sentence s compared to the top-k extracted-and-edited sentences in the target language space. Intuitively, we view the top-k extracted-and-edited sentences as the anchor points to locate a probable region in the target language space, and iteratively improve the source-to-target mapping via the comparative learning scheme.

Combined with the language modeling constraints as described in Equation 1, the final loss function for training the the translation model $V_{s\to t}$ is defined as:

$$\begin{aligned}\mathcal{L}_{s\to t}(\theta_{enc}, \theta_{dec}|\theta_R) =& \omega_{lm}\mathcal{L}_{lm}(\theta_{enc}, \theta_{dec}) + \\ & \omega_{com}\mathcal{L}_{com}(\theta_{enc}|\theta_R),\end{aligned} \tag{7}$$

where ω_{lm} and ω_{com} are hyper-parameters weighing the importance of the language modeling and the comparative learning.

Adversarial Objective Meanwhile, we need to learn a good evaluation network R to transform sentence embedding of the shared encoder into the comparable space. The evaluation network R is also shared by two languages to ensure a strong connection between two language spaces. Inspired by adversarial learning (Goodfellow et al., 2014), we can view our translation system as a "generator" that learns to generate a good translation with a higher similarity score than the extracted-and-edited sentences, and the evaluation network R as a "discriminator" that learns to rank the extracted-and-edited sentences (real sentences in the target language space) higher than the translated sentences. Thus, we have the following objective function for the evaluation network R:

$$\mathcal{L}_R(\theta_R) = -\mathbb{E}_{t'\in M'}(\log P(t'|s, M')). \tag{8}$$

Algorithm 1: The algorithm of our unsupervised NMT system with *extract-edit* approach.

Given two monolingual corpora, source $\mathcal{S}$ and target $\mathcal{T}$;
Initialization as in Section 2;
Language Modeling as in Section 2 to obtain the initialized translation model $V^{(0)}_{s\to t} = enc^{(0)} \circ dec^{(0)}$;
for $n \leftarrow 1$ **to** N **do**
  Given a source sentence s;
  Extract the top-k target sentences as the set M;
  Edit the sentences in M to obtain the set M';
  Update the evaluation network $R : \theta_R \leftarrow \text{argmin}\mathcal{L}_R$;
  Update the shared encoder and decoder R: $\theta_{enc}, \theta_{dec} \leftarrow \text{argmin}\mathcal{L}_{s\to t}$;
  Update the translation model: $V^{(n+1)}_{s\to t} = enc^{(n)} \circ dec^{(n)}$;
return $V^{(N+1)}_{s\to t} = enc^{(N)} \circ dec^{(N)}$.

Based on Equation 7 and 8, the final adversarial objective is defined as

$$\begin{aligned}&\min_{\theta_{enc},\theta_{dec}} \max_{\theta_R} \mathcal{L}(\theta_{enc}, \theta_{dec}, \theta_R) \\ &= -\mathcal{L}_R(\theta_R) + \mathcal{L}_{s\to t}(\theta_{enc}, \theta_{dec}|\theta_R),\end{aligned} \tag{9}$$

where the translation model $V_{s\to t}$ and the evaluation network R play the two-player mini-max game. We evenly alternately update between the encoder-decoder translation model and the evaluation network. The detailed training procedure is described in Algorithm 1.

3.5 Model Selection

In the fully unsupervised setting, we do not have access to parallel sentence pairs. Thus, we need to find a criterion correlated with the translation quality to select hyper-parameters. For a neural translation model $V_{s\to t}$, we propose the following criterion $D_{s\to t}$ to tune the hyper-parameters:

$$D_{s\to t} = \mathbb{E}_{s\in\mathcal{S}}[\mathbb{E}(\log P(t^*|s, M'))], \tag{10}$$

where $t^* = V_{s\to t}(s)$. Basically, we choose the hyper-parameters with the maximum expectation of the ranking scores of all translated sentences.

4 Experiments

4.1 Datasets

We consider four language pairs: English-French (*en-fr*), English-German (*en-de*), English-Russian (*en-ru*) and English-Romanian (*en-ro*) for evaluation. We use the same corpora as in Lample et al. (2018b) for these languages for fair comparison. For English, French, German and Russian, all the available sentences are used from

Model	*en*→*fr*	*fr*→*en*	*en*→*de*	*de*→*en*
LSTM Cell				
Lample et al. (2018b)	24.28	23.74	14.71	19.60
Ours (Top-1 Extract)	24.43 (+0.15)	23.90 (+0.16)	14.54 (−0.17)	19.49 (−0.11)
Ours (Top-1 Extract + Edit)	24.54 (+0.26)	24.08 (+0.34)	14.63 (−0.08)	19.57 (−0.03)
Ours (Top-10 Extract)	26.12 (+1.84)	25.83 (+2.09)	17.01 (+2.30)	21.40 (+1.80)
Ours (Top-10 Extract + Edit)	**26.97** (+2.69)	**26.66** (+2.92)	**17.48** (+2.77)	**21.93** (+2.33)
Transformer Cell				
Lample et al. (2018b)	25.14	24.18	17.16	21.00
Ours (Top-1 Extract)	25.30 (+0.16)	24.23 (+0.05)	17.12 (−0.04)	21.06 (+0.06)
Ours (Top-1 Extract + Edit)	25.44 (+0.30)	24.36 (+0.18)	17.14 (−0.02)	21.10 (+0.10)
Ours (Top-10 Extract)	26.91 (+1.77)	25.64 (+1.46)	19.11 (+1.95)	22.84 (+1.84)
Ours (Top-10 Extract + Edit)	**27.56** (+2.42)	**26.90** (+2.72)	**19.55** (+2.39)	**23.29** (+2.29)
Model	***en*→*ro***	***ro*→*en***	***en*→*ru***	***ru*→*en***
LSTM Cell				
Lample et al. (2018b)	19.65	18.52	6.24	7.83
Ours (Top-1 Extract)	19.73 (+0.08)	18.56 (+0.04)	6.32 (+0.08)	7.99 (+0.16)
Ours (Top-1 Extract + Edit)	19.81 (+0.16)	18.69 (+0.17)	6.44 (+0.20)	8.12 (+0.29)
Ours (Top-10 Extract)	21.57 (+1.92)	20.32 (+1.80)	8.87 (+2.63)	9.76 (+1.93)
Ours (Top-10 Extract + Edit)	**22.08** (+2.43)	**20.83** (+2.31)	**9.35** (+3.11)	**10.21** (+2.38)
Transformer Cell				
Lample et al. (2018b)	21.18	19.44	7.98	9.09
Ours (Top-1 Extract)	21.15 (−0.03)	19.52 (+0.08)	8.03 (+0.05)	9.20 (+0.11)
Ours (Top-1 Extract + Edit)	21.23 (+0.05)	19.59 (+0.15)	8.16 (+0.18)	9.28 (+0.19)
Ours (Top-10 Extract)	23.04 (+1.86)	21.43 (+1.99)	10.24 (+2.26)	12.29 (+3.20)
Ours (Top-10 Extract + Edit)	**23.31** (+2.13)	**21.60** (+2.16)	**11.07** (+3.09)	**12.72** (+3.63)

Table 1: The experimental results on all four language pairs and directions. The results are evaluated with BLEU metric on *newstest* 2014 for *en*↔*fr*and *newstest* 2016 for *en*↔*de*, *en*↔*ro* and *en*↔*ru*. The (+) and (−) stand for performance gains and loss separately compared with baseline models with the same NMT cells.

the WMT monolingual News Crawl datasets from years 2007 through 2017. As for Romanian, we combine the News Crawl dataset and WMT'16 monolingual dataset. The translation results are evaluated on newstest 2014 for *en-fr*, and newstest 2016 for *en-de*, *en-ro* and *en-ru*.

4.2 Implementation Details

We follow previous methods (Koehn et al., 2007; Lample et al., 2018b) to initialize our models.

Initialization We use Moses scripts (Koehn et al., 2007) for tokenization. While the system requires cross-lingual BPE embeddings to initialize the shared lookup table for related languages, we set the number of BPE codes as $60,000$. Following the previous preprocessing protocol (Lample et al., 2018b), the embeddings are then generated using fastText (Bojanowski et al., 2017) with an embedding dimension of 512, a context window of size 5 and 10 negative samples.

Model Structure In this work, the NMT models can be built upon long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) cells. For LSTM cells, both the encoder and decoder have 3 layers. As for Transformer, we use 4 layers both in the encoder and the decoder. As for both LSTM and Transformer, all encoder parameters are shared across two languages. Similarly, we share all decoder parameters across two languages. Both two model structure are optimized using Adam (Kingma and Ba, 2014) with a batch size of 32. The rate for LSTM cell is 0.0003 while Transformer's is set as 0.0001. The weights in Equation 7 are $\omega_{lm} = \omega_{ext} = 1$. The λ for calculating ranking scores is 0.5. As for the evaluation network R, we use a multilayer perceptron with two hidden layers of size 512. For efficient nearest neighbor search in the extracting step, we use the open-source Faiss library (Johnson et al., 2017)[2]. We calculate the similarity of sentences in each episode instead of each batch for computational efficiency. At decoding time, sentences are

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

generated using greedy decoding.

4.3 Results and Analysis

In this study, we aim to validate the effectiveness of extract-edit versus back-translation for unsupervised neural machine translation (NMT), so we set the unsupervised NMT method in Lample et al. (2018b) as the baseline because it currently achieves the state-of-the-art performance on all language pairs.[3] The overall translation results across four language pairs are shown in Table 1. In most of the cases, our proposed extract-edit approach can outperform the baseline models trained with back-translation. Our full models (LSTM/Transformer + Top-10 Extract + Edit) achieve more than 2 BLEU points improvement consistently across all the language pairs. Especially, on the *ru* $\rightarrow$ *en* translation with the Transformer cell, our full model surpasses the baseline score by 3.63 BLEU points. These results validate the effectiveness of our approach and indicate that the proposed extract-edit learning framework can learn a better mapping and alignment between language spaces than back-translation.

However, if extracting only top-1 target sentence in our approach, the performances are not always improved (e.g., *en* $\rightarrow$ *de*, *de* $\rightarrow$ *en*, and *en* $\rightarrow$ *ro*). Besides, *Top-10 Extract + Edit* models consistently outperforms *Top-1 Extract + Edit*. This is because more extracted-and-edited sentences lead to a higher recall, so more useful information will be used to guarantee the translation quality. The comparative translation loss can avoid the model suffering from the noise while taking advantage of more information. In other words, it is more likely to project the source sentence into the probable region in the target language space with more sentences serving as the anchor points, and the comparative learning scheme iteratively approximates towards more accurate target points. This highlights the importance of the extraction number k, which we further discuss next.

4.4 Ablation Study

The Effect of Extraction Number k As shown in Table 1, the number k of the extracted-and-edited sentences plays a vital role in our approach.

[3]Note that for a fair comparison, we are not including the results of the unsupervised phrase-based statistical machine translation system (PBSMT). But theoretically, our extract-edit learning framework can be generalized to other types of machine translation systems such as PBSMT.

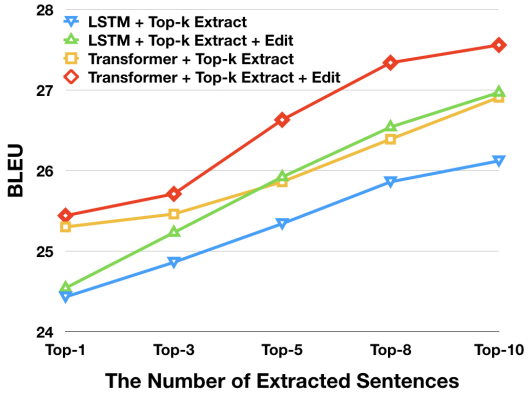

Figure 3: The effect of the number k of the extracted sentences in our approach on *en*→*fr* translation.

Thus for a more intuitive overview of its impact, we further train and evaluate multiple models with $k = 1, 3, 5, 8, 10$ on *en* $\rightarrow$ *fr* translation task. The detailed results are shown in Figure 3. This study shows that the translation performance of our approach is indeed improving as k increases. Because as we analyze above, larger k ensure a higher recall and thus more critical semantic information can be utilized to assist the translation. Besides, the diversity of extracted-and-edited sentences can potentially provide a more accurate localization of the probable region where the target sentence should be. Although we can infer that the models would perform even better with $k > 10$ from Figure 3, more computational resources will be required for that and we already observe a decelerated growth of BLEU scores from $k = 8$ to $k = 10$. Therefore, in this paper, we set $k = 10$ for the full models.

The Quality of Extraction Model In this section, we quantitatively evaluate the unsupervised extraction part of our model and compare it with the state-of-the-art supervised extraction model. Following Grégoire and Langlais (2018), we train a fully supervised parallel pair extraction model, where two Bi-LSTMs are implemented to encode sentences of two languages, and a feed-forward network is followed to culminate in a sigmoid output layer. The model is trained with around $500,000$ English-French parallel sentence pairs sampled from Europarl corpus (Koehn, 2005). As for our unsupervised extraction model, we directly use the jointly trained extraction part in our framework to extract the potential parallel sentences based on the scores computed by Equation 4. For evaluation, we sample $1,000$ parallel sentences

Noise	Model	Hits@1	Hits@3	Hits@5	Hits@8	Hits@10	Hits@15	Hits@20
0%	Supervised (Upperbound)	67.3	80.7	89.9	94.5	97.1	98.7	99.3
	Unsupervised (Ours)	52.2	54.6	68.8	80.2	89.1	91.8	93.3
50%	Supervised (Upperbound)	64.8	78.0	86.8	91.3	95.6	97.4	99.0
	Unsupervised (Ours)	46.9	49.7	62.1	73.4	83.2	87.6	89.2
90%	Supervised (Upperbound)	63.7	76.4	84.2	89.1	93.8	96.5	98.1
	Unsupervised (Ours)	41.5	46.8	58.0	69.3	77.2	83.9	87.8

Table 2: The experimental results of parallel sentence mining on the *newstest* 2012 *en* $\rightarrow$ *fr* translation dataset with different levels of added sentence noises. Metric: The percentage of Hits@k.

Cell	Learning	BLEU
LSTM	MLE Loss	12.40
	Comparative Loss	24.54
Transformer	MLE Loss	14.15
	Comparative Loss	25.44

Table 3: The performance of the unsupervised NMT systems with different learning objectives on *en* $\rightarrow$ *fr* *newstest* 2014.

from the *newstest* 2014 corpus and create three test sets with a noise ratio 0%, 50%, and 90% to simulate noisy real-world data. We report Hits@k results, which shows the percentage of the golden parallel sentences appear within the top-k place.

The detailed results are shown in Table 2. Although our extraction model structure is different from the supervised extraction model, it can still give us a good insight into the upperbound and gap of performance. We can observe a noticeable gap between unsupervised and supervised methods, but the gap is narrowing as the rank increases. Meanwhile, in our unsupervised method, the performance grows quickly when $k \leq 10$. From Table 2 we also notice that $k = 10$ is a sweet point, where the accuracy is high and the computational cost is relatively acceptable.

The Effect of Comparative Translation Finally, we aim to roughly evaluate the effect of the proposed comparable translation loss in our model. Thus, we compare our model with a two-staged NMT system, where we extract and edit the parallel pairs and retrain the NMT system with the standard maximum likelihood estimation (MLE) loss in a supervised way (by taking the extracted-and-edited sentences as the ground-truth targets). We compare the performance on the *en* $\rightarrow$ *fr* *newstest* 2014 dataset, and the results are shown in Table 3. We can observe that with the MLE loss, the translation performance will drop nearly 50%. The results indirectly reflect that the NMT systems are sensitive to noises in the training datasets. Meanwhile, it demonstrates by treating extracted-edit sentences as pivotal points instead of ground truth, our proposed comparative translation loss can avoid the NMT model suffers from the noise.

4.5 Discussion

Although our extract-edit approach can achieve better performance than the back-translation mechanism, it is still worth mentioning that our approach has more strict constraints on the domains of the source and target corpus. The extract-edit approach will work well when there is information overlap in the two language spaces. When there is little overlap in terms of domains, it will be much harder to find a good cluster of initial candidates, which may also complicate the editing process. As for the back-translation mechanism, it requires less overlap in terms of the language spaces because the language priors can be learnt in any domains. However, the corpus with matching domains can be easily obtained nowadays (e.g., Wikipedia and the news articles), which makes our extract-edit approach still widely applicable.

5 Related Work

Unsupervised NMT The current NMT systems (Sutskever et al., 2014; Cho et al., 2014a; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) are known to easily overfit and result in an inferior performance when the training data is limited (Koehn and Knowles, 2017; Isabelle et al., 2017; Sennrich, 2017). Many research efforts have been spent on how to utilize the monolingual data to improve the NMT system when only limited supervision is available (Gulcehre et al., 2015; Sennrich et al., 2016a; He et al., 2016; Zhang and Zong, 2016; Yang et al.,

2018). Recently, Lample et al. (2018a); Artetxe et al. (2018); Lample et al. (2018b) make encouraging progress on unsupervised NMT structure mainly based on initialization, denoising language modeling, and back-translation. However, all these unsupervised models are based on the back-translation learning framework to generate pseudo language pairs for training. Our work leverages the information from real target language sentences.

Comparable Corpora Mining Comparable corpora mining aims at extracting parallel sentences from comparable monolingual corpora such as news stories written on the same topic in different languages. Most of the previous methods align the documents based on metadata and then extract parallel sentences using human-defined features (Munteanu and Marcu, 2002, 2006; Hewavitharana and Vogel, 2011). Recent neural-based methods (Chu et al., 2016; Grover and Mitra, 2017; Grégoire and Langlais, 2018) learn to identify parallel sentences in the semantic spaces. However, these methods require large amounts of parallel sentence pairs to train the systems first and then test the performance on raw comparable corpora, which does not apply to languages with limited resources. Instead, we explore the corpora mining in an unsupervised fashion and propose a joint training framework with machine translation.

Retrieval-Augmented Text Generation Our work is also related to the recent work on applying retrieval mechanisms to augment text generation, such as image captioning (Kuznetsova et al., 2013; Mason and Charniak, 2014), dialogue generation (Song et al., 2016; Yan et al., 2016; Wu et al., 2018) and style transfer (Lin et al., 2017; Li et al., 2018). Some editing-based models (Guu et al., 2018; Hashimoto et al., 2018) are proposed to further enhance the retrieved text. Recent work in machine translation (Gu et al., 2018) augments an NMT model with sentence pairs retrieved by an off-the-shelf search engine. However, these methods are two-staged with supervised retrieval first. In our work, the extracted-edited sentences are not directly used as the ground truth to train the translation model. Instead, we view these sentences as pivotal points in the target language space and further we propose a comparative translation loss to train the system in a fully unsupervised way.

6 Conclusion

In this paper, we propose an *extract-edit* approach, an effective alternative to the widely-used back-translation in unsupervised NMT. Instead of generating pseudo language pairs to train the systems with the reconstruction loss, we design a comparative translation loss that leverages real sentences in the target language space. Empirically, our method advances the previous state-of-the-art NMT systems across four language pairs using the monolingual corpora only. Theoretically, we believe the extract-edit learning framework can be generalized to other types of unsupervised machine translation systems and even some other unsupervised learning tasks.

Acknowledgments

The authors would like to thank the anonymous reviewers for their thoughtful comments. The work was supported by the Facebook Low Resource Neural Machine Translation Research Award. The authors are solely responsible for the contents of the paper, and the opinions expressed in this publication do not reflect those of the funding agencies.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *Proceedings of the 4th International Conference for Learning Representations (ICLR)*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics (TACL)*, 5(1):135–146.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2016. Parallel sentence extraction from comparable corpora with neural network features. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1243–1252.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of 28th Conference on neural information processing systems (NeurIPS)*, pages 2672–2680.

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 (COLING)*.

Jeenu Grover and Pabitra Mitra. 2017. Bilingual word embeddings with bucketed cnn for parallel sentence extraction. In *Proceedings of ACL 2017, Student Research Workshop*, pages 11–16.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided non-parametric neural machine translation. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI)*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association of Computational Linguistics (TACL)*, 6:437–450.

Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy S Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. In *Proceedings of 32th Conference on Neural Information Processing Systems (NeurIPS)*, pages 10073–10083.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 820–828.

Sanjika Hewavitharana and Stephan Vogel. 2011. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 61–68.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2486–2496.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2013. Generalizing image captions for image-text parallel corpus. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 790–796.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

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

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1865–1874.

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In *Proceedings of the 31th Conference on Neural Information Processing Systems (NeurIP)*, pages 3155–3165.

Rebecca Mason and Eugene Charniak. 2014. Domain-specific image captioning. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*, pages 11–20.

Dragos Stefan Munteanu and Daniel Marcu. 2002. Processing comparable corpora with bilingual suffix trees. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 81–88.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 376–382.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96.

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

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of Advances in neural information processing systems (NeurIPS)*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of 31st Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Yu Wu, Furu Wei, Shaohan Huang, Zhoujun Li, and Ming Zhou. 2018. Response generation by context-aware prototype editing. *arXiv preprint arXiv:1806.07042*.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, pages 55–64.

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 (ACL)*, pages 46–55.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1535–1545.