Semi-Supervised Learning for Neural Machine Translation

Yong Cheng[#], Wei Xu[#], Zhongjun He[+], Wei He[+], Hua Wu[+], Maosong Sun[†] and Yang Liu[†] *
[#]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[†]State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[+]Baidu Inc., Beijing, China
chengyong3001@gmail.com weixu@tsinghua.edu.cn
{hezhongjun,hewei06,wu_hua}@baidu.com
{sms,liuyang2011}@tsinghua.edu.cn

Abstract

While end-to-end neural machine translation (NMT) has made remarkable progress recently, NMT systems only rely on parallel corpora for parameter estimation. Since parallel corpora are usually limited in quantity, quality, and coverage, especially for low-resource languages, it is appealing to exploit monolingual corpora to improve NMT. We propose a semi-supervised approach for training NMT models on the concatenation of labeled (parallel corpora) and unlabeled (monolingual corpora) data. The central idea is to reconstruct the monolingual corpora using an autoencoder, in which the source-to-target and target-to-source translation models serve as the encoder and decoder, respectively. Our approach can not only exploit the monolingual corpora of the target language, but also of the source language. Experiments on the Chinese-English dataset show that our approach achieves significant improvements over state-of-the-art SMT and NMT systems.

1 Introduction

End-to-end neural machine translation (NMT), which leverages a single, large neural network to directly transform a source-language sentence into a target-language sentence, has attracted increasing attention in recent several years (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015). Free of latent structure design and feature engineering that are critical in conventional statistical machine translation (SMT) (Brown et al., 1993; Koehn et al., 2003; Chiang, 2005), NMT has proven to excel in modeling long-distance dependencies by enhancing recurrent neural networks (RNNs) with the gating (Hochreiter and Schmidhuber, 1993; Cho et al., 2014; Sutskever et al., 2014) and attention mechanisms (Bahdanau et al., 2015).

However, most existing NMT approaches suffer from a major drawback: they heavily rely on parallel corpora for training translation models. This is because NMT directly models the probability of a target-language sentence given a source-language sentence and does not have a separate language model like SMT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015). Unfortunately, parallel corpora are usually only available for a handful of resource-rich languages and restricted to limited domains such as government documents and news reports. In contrast, SMT is capable of exploiting abundant target-side monolingual corpora to boost fluency of translations. Therefore, the unavailability of large-scale, high-quality, and wide-coverage parallel corpora hinders the applicability of NMT.

As a result, several authors have tried to use abundant monolingual corpora to improve NMT. Gulccehre et al. (2015) propose two methods, which are referred to as shallow fusion and deep fusion, to integrate a language model into NMT. The basic idea is to use the language model to score the candidate words proposed by the translation model at each time step or concatenating the hidden states of the language model and the decoder. Although their approach leads to significant improvements, one possible downside is that the network architecture has to be modified to integrate the language model.

Alternatively, Sennrich et al. (2015) propose two approaches to exploiting monolingual corpora that is transparent to network architectures. The first approach pairs monolingual sentences with dummy input. Then, the parameters of encoder

* Yang Liu is the corresponding author.

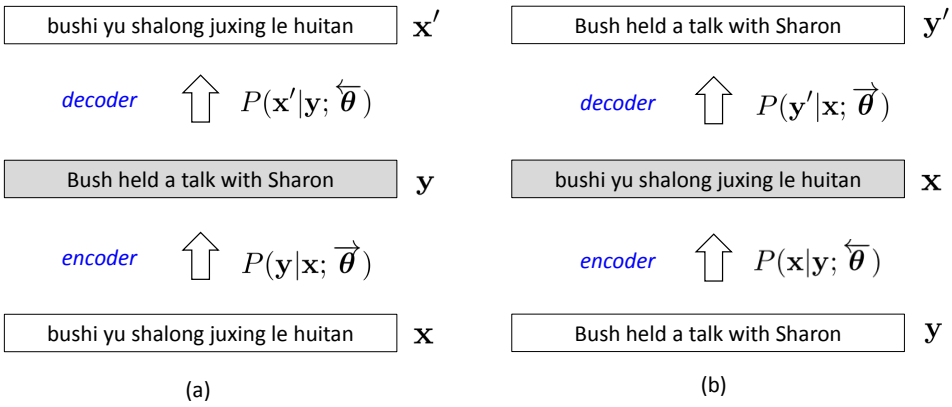

Figure 1: Examples of (a) source autoencoder and (b) target autoencoder on monolingual corpora. Our idea is to leverage autoencoders to exploit monolingual corpora for NMT. In a source autoencoder, the source-to-target model $P(\mathbf{y}|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})$ serves as an encoder to transform the observed source sentence $\mathbf{x}$ into a latent target sentence $\mathbf{y}$ (highlighted in grey), from which the target-to-source model $P(\mathbf{x}'|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})$ reconstructs a copy of the observed source sentence $\mathbf{x}'$ from the latent target sentence. As a result, monolingual corpora can be combined with parallel corpora to train bidirectional NMT models in a semi-supervised setting.

and attention model are fixed when training on these pseudo parallel sentence pairs. In the second approach, they first train a nerual translation model on the parallel corpus and then use the learned model to translate a monolingual corpus. The monolingual corpus and its translations constitute an additional pseudo parallel corpus. Similar ideas have also been suggested in conventional SMT (Ueffing et al., 2007; Bertoldi and Federico, 2009). Sennrich et al. (2015) report that their approach significantly improves translation quality across a variety of language pairs.

In this paper, we propose semi-supervised learning for neural machine translation. Given labeled (i.e., parallel corpora) and unlabeled (i.e., monolingual corpora) data, our approach jointly trains source-to-target and target-to-source translation models. The key idea is to append a reconstruction term to the training objective, which aims to reconstruct the observed monolingual corpora using an autoencoder. In the autoencoder, the source-to-target and target-to-source models serve as the encoder and decoder, respectively. As the inference is intractable, we propose to sample the full search space to improve the efficiency. Specifically, our approach has the following advantages:

1. *Transparent to network architectures*: our approach does not depend on specific architectures and can be easily applied to arbitrary end-to-end NMT systems.
2. *Both the source and target monolingual corpora can be used*: our approach can benefit NMT not only using target monolingual corpora in a conventional way, but also the monolingual corpora of the source language.

Experiments on Chinese-English NIST datasets show that our approach results in significant improvements in both directions over state-of-the-art SMT and NMT systems.

2 Semi-Supervised Learning for Neural Machine Translation

2.1 Supervised Learning

Given a parallel corpus $\mathcal{D} = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$, the standard training objective in NMT is to maximize the likelihood of the training data:

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta}), \quad (1)$$

where $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ is a neural translation model and $\boldsymbol{\theta}$ is a set of model parameters. $\mathcal{D}$ can be seen as *labeled* data for the task of predicting a target sentence $\mathbf{y}$ given a source sentence $\mathbf{x}$.

As $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ is modeled by a single, large neural network, there does not exist a separate target language model $P(\mathbf{y};\boldsymbol{\theta})$ in NMT. Therefore, parallel corpora have been the only resource for parameter estimation in most existing NMT systems. Unfortunately, even for a handful of resource-rich

languages, the available domains are unbalanced and restricted to government documents and news reports. Therefore, the availability of large-scale, high-quality, and wide-coverage parallel corpora becomes a major obstacle for NMT.

2.2 Autoencoders on Monolingual Corpora

It is appealing to explore the more readily available, abundant monolingual corpora to improve NMT. Let us first consider an *unsupervised* setting: how to train NMT models on a monolingual corpus $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$?

Our idea is to leverage *autoencoders* (Vincent et al., 2010; Socher et al., 2011): (1) *encoding* an observed target sentence into a latent source sentence using a target-to-source translation model and (2) *decoding* the source sentence to reconstruct the observed target sentence using a source-to-target model. For example, as shown in Figure 1(b), given an observed English sentence "Bush held a talk with Sharon", a target-to-source translation model (i.e., encoder) transforms it into a Chinese translation "bushi yu shalong juxing le huitan" that is unobserved on the training data (highlighted in grey). Then, a source-to-target translation model (i.e., decoder) reconstructs the observed English sentence from the Chinese translation.

More formally, let $P(\mathbf{y}|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})$ and $P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})$ be *source-to-target* and *target-to-source* translation models respectively, where $\overrightarrow{\boldsymbol{\theta}}$ and $\overleftarrow{\boldsymbol{\theta}}$ are corresponding model parameters. An autoencoder aims to reconstruct the observed target sentence via a latent source sentence:

$$\begin{aligned} & P(\mathbf{y}'|\mathbf{y};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}}) \\ = & \sum_{\mathbf{x}} P(\mathbf{y}',\mathbf{x}|\mathbf{y};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}}) \\ = & \sum_{\mathbf{x}} \underbrace{P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})}_{encoder} \underbrace{P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}_{decoder}, \end{aligned} \quad (2)$$

where $\mathbf{y}$ is an observed target sentence, $\mathbf{y}'$ is a copy of $\mathbf{y}$ to be reconstructed, and $\mathbf{x}$ is a latent source sentence.

We refer to Eq. (2) as a *target autoencoder*. [1] Likewise, given a monolingual corpus of source language $\mathcal{S} = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$, it is natural to introduce a *source autoencoder* that aims at reconstructing the observed source sentence via a latent target sentence:

$$\begin{aligned} & P(\mathbf{x}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}}) \\ = & \sum_{\mathbf{y}} P(\mathbf{x}',\mathbf{y}|\mathbf{x};\overleftarrow{\boldsymbol{\theta}}) \\ = & \sum_{\mathbf{y}} \underbrace{P(\mathbf{y}|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}_{encoder} \underbrace{P(\mathbf{x}'|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})}_{decoder}. \end{aligned} \quad (3)$$

Please see Figure 1(a) for illustration.

2.3 Semi-Supervised Learning

As the autoencoders involve both source-to-target and target-to-source models, it is natural to combine parallel corpora and monolingual corpora to learn birectional NMT translation models in a semi-supervised setting.

Formally, given a parallel corpus $\mathcal{D} = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$, a monolingual corpus of target language $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$, and a monolingual corpus of source language $\mathcal{S} = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$, we introduce our new semi-supervised training objective as follows:

$$\begin{aligned} & J(\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}}) \\ = & \underbrace{\sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)};\overrightarrow{\boldsymbol{\theta}})}_{\textit{source-to-target likelihood}} \\ & + \underbrace{\sum_{n=1}^{N} \log P(\mathbf{x}^{(n)}|\mathbf{y}^{(n)};\overleftarrow{\boldsymbol{\theta}})}_{\textit{target-to-source likelihood}} \\ & + \lambda_1 \underbrace{\sum_{t=1}^{T} \log P(\mathbf{y}'|\mathbf{y}^{(t)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})}_{\textit{target autoencoder}} \\ & + \lambda_2 \underbrace{\sum_{s=1}^{S} \log P(\mathbf{x}'|\mathbf{x}^{(s)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})}_{\textit{source autoencoder}}, \end{aligned} \quad (4)$$

where λ_1 and λ_2 are hyper-parameters for balancing the preference between likelihood and autoencoders.

Note that the objective consists of four parts: source-to-target likelihood, target-to-source likelihood, target autoencoder, and source autoencoder. In this way, our approach is capable of exploiting abundant monolingual corpora of both source and target languages.

[1] Our definition of auotoencoders is inspired by Ammar et al. (2014). Note that our autoencoders inherit the same spirit from conventional autoencoders (Vincent et al., 2010; Socher et al., 2011) except that the hidden layer is denoted by a latent sentence instead of real-valued vectors.

The optimal model parameters are given by

$$\overrightarrow{\boldsymbol{\theta}}^* = \text{argmax}\left\{\sum_{n=1}^{N}\log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)};\overrightarrow{\boldsymbol{\theta}}) + \lambda_1\sum_{t=1}^{T}\log P(\mathbf{y}'|\mathbf{y}^{(t)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}}) + \lambda_2\sum_{s=1}^{S}\log P(\mathbf{x}'|\mathbf{x}^{(s)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})\right\} \quad (5)$$

$$\overleftarrow{\boldsymbol{\theta}}^* = \text{argmax}\left\{\sum_{n=1}^{N}\log P(\mathbf{x}^{(n)}|\mathbf{y}^{(n)};\overleftarrow{\boldsymbol{\theta}}) + \lambda_1\sum_{t=1}^{T}\log P(\mathbf{y}'|\mathbf{y}^{(t)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}}) + \lambda_2\sum_{s=1}^{S}\log P(\mathbf{x}'|\mathbf{x}^{(s)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})\right\} \quad (6)$$

It is clear that the source-to-target and target-to-source models are connected via the autoencoder and can hopefully benefit each other in joint training.

2.4 Training

We use mini-batch stochastic gradient descent to train our joint model. For each iteration, besides the mini-batch from the parallel corpus, we also construct two additional mini-batches by randomly selecting sentences from the source and target monolingual corpora. Then, gradients are collected from these mini-batches to update model parameters.

The partial derivative of $J(\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})$ with respect to the source-to-target model $\overrightarrow{\boldsymbol{\theta}}$ is given by

$$\begin{aligned}&\frac{\partial J(\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})}{\partial\overrightarrow{\boldsymbol{\theta}}}\\ =&\sum_{n=1}^{N}\frac{\partial\log P(\mathbf{y}^{(n)}|\mathbf{x}^{(n)};\overrightarrow{\boldsymbol{\theta}})}{\partial\overrightarrow{\boldsymbol{\theta}}}\\ &+\lambda_1\sum_{t=1}^{T}\frac{\partial\log P(\mathbf{y}'|\mathbf{y}^{(t)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})}{\partial\overrightarrow{\boldsymbol{\theta}}}\\ &+\lambda_2\sum_{s=1}^{S}\frac{\partial\log P(\mathbf{x}'|\mathbf{x}^{(s)};\overrightarrow{\boldsymbol{\theta}},\overleftarrow{\boldsymbol{\theta}})}{\partial\overrightarrow{\boldsymbol{\theta}}}. \quad (7)\end{aligned}$$

The partial derivative with respect to $\overleftarrow{\boldsymbol{\theta}}$ can be calculated similarly.

Unfortunately, the second and third terms in Eq. (7) are intractable to calculate due to the exponential search space. For example, the derivative in the third term in Eq. (7) is given by

$$\frac{\sum_{\mathbf{x}\in\mathcal{X}(\mathbf{y})}P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})\frac{\partial\log P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}{\partial\overrightarrow{\boldsymbol{\theta}}}}{\sum_{\mathbf{x}\in\mathcal{X}(\mathbf{y})}P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}. \quad (8)$$

It is prohibitively expensive to compute the sums due to the exponential search space of $\mathcal{X}(\mathbf{y})$.

Alternatively, we propose to use a subset of the full space $\tilde{\mathcal{X}}(\mathbf{y}) \subset \mathcal{X}(\mathbf{y})$ to approximate Eq. (8):

$$\frac{\sum_{\mathbf{x}\in\tilde{\mathcal{X}}(\mathbf{y})}P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})\frac{\partial\log P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}{\partial\overrightarrow{\boldsymbol{\theta}}}}{\sum_{\mathbf{x}\in\tilde{\mathcal{X}}(\mathbf{y})}P(\mathbf{x}|\mathbf{y};\overleftarrow{\boldsymbol{\theta}})P(\mathbf{y}'|\mathbf{x};\overrightarrow{\boldsymbol{\theta}})}. \quad (9)$$

In practice, we use the top-k list of candidate translations of $\mathbf{y}$ as $\tilde{\mathcal{X}}(\mathbf{y})$. As $|\tilde{\mathcal{X}}(\mathbf{y})| \ll \mathcal{X}|(\mathbf{y})|$, it is possible to calculate Eq. (9) efficiently by enumerating all candidates in $\tilde{\mathcal{X}}(\mathbf{y})$. In practice, we find this approximation results in significant improvements and $k = 10$ seems to suffice to keep the balance between efficiency and translation quality.

		Chinese	English
Parallel	# Sent.	2.56M	
	# Word	67.54M	74.82M
	Vocab.	0.21M	0.16M
Monolingual	# Sent.	18.75M	22.32M
	# Word	451.94M	399.83M
	Vocab.	0.97M	1.34M

Table 1: Characteristics of parallel and monolingual corpora.

3 Experiments

3.1 Setup

We evaluated our approach on the Chinese-English dataset.

As shown in Table 1, we use both a parallel corpus and two monolingual corpora as the training set. The parallel corpus from LDC consists of 2.56M sentence pairs with 67.53M Chinese words and 74.81M English words. The vocabulary sizes of Chinese and English are 0.21M and 0.16M, respectively. We use the Chinese and English parts of the Xinhua portion of the GIGAWORD corpus as the monolingual corpora. The Chinese monolingual corpus contains 18.75M sentences with 451.94M words. The English corpus contains 22.32M sentences with 399.83M words. The vocabulary sizes of Chinese and English are 0.97M and 1.34M, respectively.

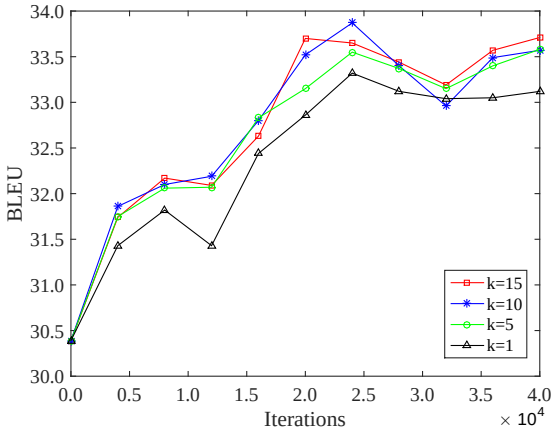

Figure 2: Effect of sample size k on the Chinese-to-English validation set.

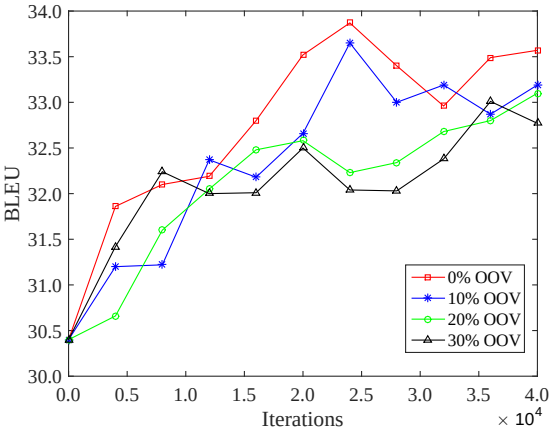

Figure 4: Effect of OOV ratio on the Chinese-to-English validation set.

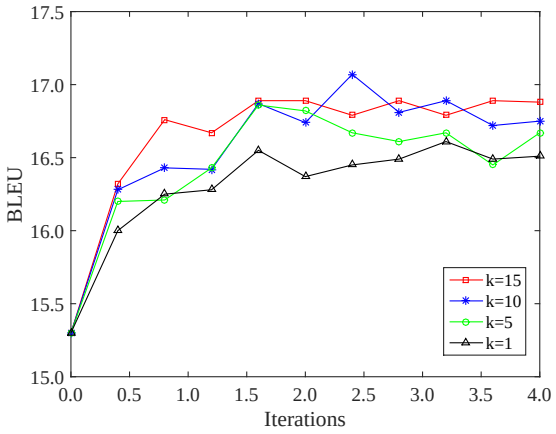

Figure 3: Effect of sample size k on the English-to-Chinese validation set.

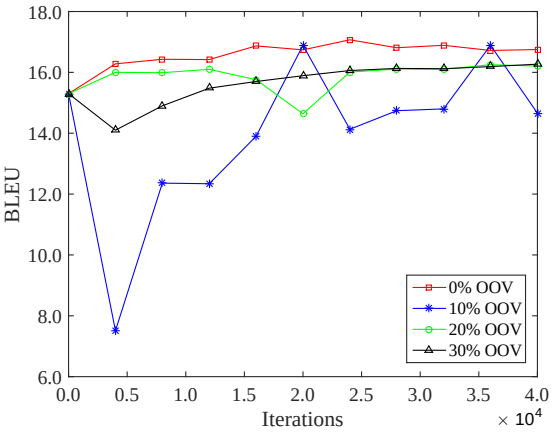

Figure 5: Effect of OOV ratio on the English-to-Chinese validation set.

For Chinese-to-English translation, we use the NIST 2006 Chinese-English dataset as the validation set for hyper-parameter optimization and model selection. The NIST 2002, 2003, 2004, and 2005 datasets serve as test sets. Each Chinese sentence has four reference translations. For English-to-Chinese translation, we use the NIST datasets in a reverse direction: treating the first English sentence in the four reference translations as a source sentence and the original input Chinese sentence as the single reference translation. The evaluation metric is case-insensitive BLEU (Papineni et al., 2002) as calculated by the `multi-bleu.perl` script.

We compared our approach with two state-of-the-art SMT and NMT systems:

1. MOSES (Koehn et al., 2007): a phrase-based SMT system;
2. RNNSEARCH (Bahdanau et al., 2015): an attention-based NMT system.

For MOSES, we use the default setting to train the phrase-based translation on the parallel corpus and optimize the parameters of log-linear models using the minimum error rate training algorithm (Och, 2003). We use the SRILM toolkit (Stolcke, 2002) to train 4-gram language models.

For RNNSEARCH, we use the parallel corpus to train the attention-based neural translation models. We set the vocabulary size of word embeddings to 30K for both Chinese and English. We follow Luong et al. (2015) to address rare words.

On top of RNNSEARCH, our approach is capable of training bidirectional attention-based neural translation models on the concatenation of parallel and monolingual corpora. The sample size k is set to 10. We set the hyper-parameter $\lambda_1 = 0.1$ and

$\lambda_2 = 0$ when we add the target monolingual corpus, and $\lambda_1 = 0$ and $\lambda_2 = 0.1$ for source monolingual corpus incorporation. The threshold of gradient clipping is set to 0.05. The parameters of our model are initialized by the model trained on parallel corpus.

3.2 Effect of Sample Size k

As the inference of our approach is intractable, we propose to approximate the full search space with the top-k list of candidate translations to improve efficiency (see Eq. (9)).

Figure 2 shows the BLEU scores of various settings of k over time. Only the English monolingual corpus is appended to the training data. We observe that increasing the size of the approximate search space generally leads to improved BLEU scores. There are significant gaps between $k = 1$ and $k = 5$. However, keeping increasing k does not result in significant improvements and decreases the training efficiency. We find that $k = 10$ achieves a balance between training efficiency and translation quality. As shown in Figure 3, similar findings are also observed on the English-to-Chinese validation set. Therefore, we set $k = 10$ in the following experiments.

3.3 Effect of OOV Ratio

Given a parallel corpus, what kind of monolingual corpus is most beneficial for improving translation quality? To answer this question, we investigate the effect of *OOV ratio* on translation quality, which is defined as

$$\text{ratio} = \frac{\sum_{y \in \mathbf{y}} [\![y \notin \mathcal{V}_{D_t}]\!]}{|\mathbf{y}|}, \quad (10)$$

where $\mathbf{y}$ is a target-language sentence in the monolingual corpus $\mathcal{T}$, y is a target-language word in $\mathbf{y}$, $\mathcal{V}_{D_t}$ is the vocabulary of the target side of the parallel corpus D.

Intuitively, the OOV ratio indicates how a sentence in the monolingual resembles the parallel corpus. If the ratio is 0, all words in the monolingual sentence also occur in the parallel corpus.

Figure 4 shows the effect of OOV ratio on the Chinese-to-English validation set. Only English monolingual corpus is appended to the parallel corpus during training. We constructed four monolingual corpora of the same size in terms of sentence pairs. "0% OOV" means the OOV ratio is 0% for all sentences in the monolingual corpus. "10% OOV" suggests that the OOV ratio is no greater 10% for each sentence in the monolingual corpus. We find that using a monolingual corpus with a lower OOV ratio generally leads to higher BLEU scores. One possible reason is that low-OOV monolingual corpus is relatively easier to reconstruct than its high-OOV counterpart and results in better estimation of model parameters.

Figure 5 shows the effect of OOV ratio on the English-to-Chinese validation set. Only English monolingual corpus is appended to the parallel corpus during training. We find that "0% OOV" still achieves the highest BLEU scores.

3.4 Comparison with SMT

Table 2 shows the comparison between MOSES and our work. MOSES used the monolingual corpora as shown in Table 1: 18.75M Chinese sentences and 22.32M English sentences. We find that exploiting monolingual corpora dramatically improves translation performance in both Chinese-to-English and English-to-Chinese directions.

Relying only on parallel corpus, RNNSEARCH outperforms MOSES trained also only on parallel corpus. But the capability of making use of abundant monolingual corpora enables MOSES to achieve much higher BLEU scores than RNNSEARCH only using parallel corpus.

Instead of using all sentences in the monolingual corpora, we constructed smaller monolingual corpora with zero OOV ratio: 2.56M Chinese sentences with 47.51M words and 2.56M English English sentences with 37.47M words. In other words, the monolingual corpora we used in the experiments are much smaller than those used by MOSES.

By adding English monolingual corpus, our approach achieves substantial improvements over RNNSEARCH using only parallel corpus (up to +4.7 BLEU points). In addition, significant improvements are also obtained over MOSES using both parallel and monolingual corpora (up to +3.5 BLEU points).

An interesting finding is that adding English monolingual corpora helps to improve English-to-Chinese translation over RNNSEARCH using only parallel corpus (up to +3.2 BLEU points), suggesting that our approach is capable of improving NMT using source-side monolingual corpora.

In the English-to-Chinese direction, we obtain similar findings. In particular, adding Chi-

System	Training Data			Direction	NIST06	NIST02	NIST03	NIST04	NIST05
	CE	C	E						
MOSES	√	×	×	C → E	32.48	32.69	32.39	33.62	30.23
				E → C	14.27	18.28	15.36	13.96	14.11
	√	×	√	C → E	34.59	35.21	35.71	35.56	33.74
	√	√	×	E → C	20.69	25.85	19.76	18.77	19.74
RNNSEARCH	√	×	×	C →E	30.74	35.16	33.75	34.63	31.74
				E →C	15.71	20.76	16.56	16.85	15.14
	√	×	√	C → E	35.61^{**++}	38.78^{**++}	38.32^{**++}	38.49^{**++}	36.45^{**++}
				E →C	17.59^{++}	23.99^{++}	18.95^{++}	18.85^{++}	17.91^{++}
	√	√	×	C →E	35.01^{++}	38.20^{**++}	37.99^{**++}	38.16^{**++}	36.07^{**++}
				E →C	21.12^{*++}	29.52^{*++}	20.49^{*++}	21.59^{*++}	19.97^{++}

Table 2: Comparison with MOSES and RNNSEARCH. MOSES is a phrase-based statistical machine translation system (Koehn et al., 2007). RNNSEARCH is an attention-based neural machine translation system (Bahdanau et al., 2015). "CE" donates Chinese-English parallel corpus, "C" donates Chinese monolingual corpus, and "E" donates English monolingual corpus. "√" means the corpus is included in the training data and × means not included. "NIST06" is the validation set and "NIST02-05" are test sets. The BLEU scores are case-insensitive. "*": significantly better than MOSES ($p < 0.05$); "**": significantly better than MOSES ($p < 0.01$);"+": significantly better than RNNSEARCH ($p < 0.05$); "++": significantly better than RNNSEARCH ($p < 0.01$).

Method	Training Data			Direction	NIST06	NIST02	NIST03	NIST04	NIST05
	CE	C	E						
Sennrich et al. (2015)	√	×	√	C →E	34.10	36.95	36.80	37.99	35.33
	√	√	×	E →C	19.85	28.83	20.61	20.54	19.17
this work	√	×	√	C →E	35.61^{**}	38.78^{**}	38.32^{**}	38.49^{*}	36.45^{**}
				E →C	17.59	23.99	18.95	18.85	17.91
	√	√	×	C →E	35.01^{**}	38.20^{**}	37.99^{**}	38.16	36.07^{**}
				E →C	21.12^{**}	29.52^{**}	20.49	21.59^{**}	19.97^{**}

Table 3: Comparison with Sennrich et al. (2015). Both Sennrich et al. (2015) and our approach build on top of RNNSEARCH to exploit monolingual corpora. The BLEU scores are case-insensitive. "*": significantly better than Sennrich et al. (2015) ($p < 0.05$); "**": significantly better than Sennrich et al. (2015) ($p < 0.01$).

nese monolingual corpus leads to more benefits to English-to-Chinese translation than adding English monolingual corpus. We also tried to use both Chinese and English monolingual corpora through simply setting all the λ to 0.1 but failed to obtain further significant improvements.

Therefore, our findings can be summarized as follows:

1. Adding target monolingual corpus improves over using only parallel corpus for source-to-target translation;

2. Adding source monolingual corpus also improves over using only parallel corpus for source-to-target translation, but the improvements are smaller than adding target monolingual corpus;

3. Adding both source and target monolingual corpora does not lead to further significant improvements.

3.5 Comparison with Previous Work

We re-implemented Sennrich et al. (2015)'s method on top of RNNSEARCH as follows:

1. Train the target-to-source neural translation model $P(\mathbf{x}|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$ on the parallel corpus $D = \{\langle \mathbf{x}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$.

2. The trained target-to-source model $\overleftarrow{\boldsymbol{\theta}}^{*}$ is used to translate a target monolingual corpus $\mathcal{T} = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$ into a source monolingual corpus $\tilde{\mathcal{S}} = \{\tilde{\mathbf{x}}^{(t)}\}_{t=1}^{T}$.

3. The target monolingual corpus is paired with its translations to form a pseudo parallel corpus, which is then appended to the original parallel corpus to obtain a larger parallel corpus: $\tilde{\mathcal{D}} = \mathcal{D} \cup \langle \tilde{\mathcal{S}}, \mathcal{T} \rangle$.

4. Re-train the the source-to-target neural translation model on $\tilde{\mathcal{D}}$ to obtain the final model parameters $\overrightarrow{\boldsymbol{\theta}}^{*}$.

Monolingual	hongsen shuo , ruguo you na jia famu gongsi dangan yishenshifa , name tamen jiang zihui qiancheng .
Reference	hongsen said, if any *logging companies* dare to defy the law, then they will *destroy their own future* .
Translation	hun sen said , if any of *those companies* dare defy the law , then they will *have their own fate* . [*iteration 0*]
	hun sen said if any *tree felling company* dared to break the law , then they would *kill themselves* . [*iteration 40K*]
	hun sen said if any *logging companies* dare to defy the law , they would *destroy the future themselves* . [*iteration 240K*]

Monolingual	dan yidan panjue jieguo zuizhong queding , ze bixu zai 30 tian nei zhixing .
Reference	But once *the final verdict is confirmed* , it must be executed within 30 days .
Translation	however , *in the final analysis* , it must be carried out within 30 days . [*iteration 0*]
	however , *in the final analysis* , the final decision will be carried out within 30 days . [*iteration 40K*]
	however , once *the verdict is finally confirmed* , it must be carried out within 30 days . [*iteration 240K*]

Table 4: Example translations of sentences in the monolingual corpus during semi-supervised learning. We find our approach is capable of generating better translations of the monolingual corpus over time.

Table 3 shows the comparison results. Both the two approaches use the same parallel and monolingual corpora. Our approach achieves significant improvements over Sennrich et al. (2015) in both Chinese-to-English and English-to-Chinese directions (up to +1.8 and +1.0 BLEU points). One possible reason is that Sennrich et al. (2015) only use the pesudo parallel corpus for parameter estimation for once (see Step 4 above) while our approach enables source-to-target and target-to-source models to interact with each other iteratively on both parallel and monolingual corpora.

To some extent, our approach can be seen as an iterative extension of Sennrich et al. (2015)'s approach: after estimating model parameters on the pseudo parallel corpus, the learned model parameters are used to produce a better pseudo parallel corpus. Table 4 shows example Viterbi translations on the Chinese monolingual corpus over iterations:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \left\{ P(\mathbf{y}'|\mathbf{x}; \overrightarrow{\boldsymbol{\theta}}) P(\mathbf{x}|\mathbf{y}; \overleftarrow{\boldsymbol{\theta}}) \right\}. \quad (11)$$

We observe that the quality of Viterbi translations generally improves over time.

4 Related Work

Our work is inspired by two lines of research: (1) exploiting monolingual corpora for machine translation and (2) autoencoders in unsupervised and semi-supervised learning.

4.1 Exploiting Monolingual Corpora for Machine Translation

Exploiting monolingual corpora for conventional SMT has attracted intensive attention in recent years. Several authors have introduced transductive learning to make full use of monolingual corpora (Ueffing et al., 2007; Bertoldi and Federico, 2009). They use an existing translation model to translate unseen source text, which can be paired with its translations to form a pseudo parallel corpus. This process iterates until convergence. While Klementiev et al. (2012) propose an approach to estimating phrase translation probabilities from monolingual corpora, Zhang and Zong (2013) directly extract parallel phrases from monolingual corpora using retrieval techniques. Another important line of research is to treat translation on monolingual corpora as a decipherment problem (Ravi and Knight, 2011; Dou et al., 2014).

Closely related to Gulccehre et al. (2015) and Sennrich et al. (2015), our approach focuses on learning birectional NMT models via autoencoders on monolingual corpora. The major advantages of our approach are the transparency to network architectures and the capability to exploit both source and target monolingual corpora.

4.2 Autoencoders in Unsupervised and Semi-Supervised Learning

Autoencoders and their variants have been widely used in unsupervised deep learning ((Vincent et al., 2010; Socher et al., 2011; Ammar et al., 2014), just to name a few). Among them, Socher et al. (2011)'s approach bears close resemblance to our approach as they introduce semi-supervised recursive autoencoders for sentiment analysis. The difference is that we are interested in making a better use of parallel and monolingual corpora while they concentrate on injecting partial supervision to conventional unsupervised autoencoders. Dai and Le (2015) introduce a sequence autoencoder to reconstruct an observed sequence via RNNs. Our approach differs from sequence autoencoders in that we use bidirectional translation models as encoders and decoders to enable them to interact within the autoencoders.

5 Conclusion

We have presented a semi-supervised approach to training bidirectional neural machine translation models. The central idea is to introduce autoencoders on the monolingual corpora with source-to-target and target-to-source translation models as encoders and decoders. Experiments on Chinese-English NIST datasets show that our approach leads to significant improvements.

As our method is sensitive to the OOVs present in monolingual corpora, we plan to integrate Jean et al. (2015)'s technique on using very large vocabulary into our approach. It is also necessary to further validate the effectiveness of our approach on more language pairs and NMT architectures. Another interesting direction is to enhance the connection between source-to-target and target-to-source models (e.g., letting the two models share the same word embeddings) to help them benefit more from interacting with each other.

Acknowledgements

This work was done while Yong Cheng was visiting Baidu. This research is supported by the 973 Program (2014CB340501, 2014CB340505), the National Natural Science Foundation of China (No. 61522204, 61331013, 61361136003), 1000 Talent Plan grant, Tsinghua Initiative Research Program grants 20151080475 and a Google Faculty Research Award. We sincerely thank the viewers for their valuable suggestions.

References

Waleed Ammar, Chris Dyer, and Noah Smith. 2014. Conditional random field autoencoders for unsupervised structred prediction. In *Proceedings of NIPS 2014*.

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation. In *Proceedings of WMT*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguisitics*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*.

Kyunhyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST-8*.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Proceedings of NIPS*.

Qing Dou, Ashish Vaswani, and Kevin Knight. 2014. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In *Proceedings of EMNLP*.

Caglar Gulccehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. arXiv:1503.03535 [cs.CL].

Sepp Hochreiter and Jürgen Schmidhuber. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguisitics*.

Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of EMNLP*.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*.

Alexandre Klementiev, Ann Irvine, Chris Callison-Burch, and David Yarowsky. 2012. Toward statistical machine translation without paralel corpora. In *Proceedings of EACL*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL (demo session)*.

Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of ACL*.

Franz Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a methof for automatic evaluation of machine translation. In *Proceedings of ACL*.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving nerual machine translation models with monolingual data. arXiv:1511.06709 [cs.CL].

Richard Socher, Jeffrey Pennington, Eric Huang, Andrew Ng, and Christopher Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of EMNLP*.

Andreas Stolcke. 2002. Srilm - am extensible language modeling toolkit. In *Proceedings of ICSLP*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Trasductive learning for statistical machine translation. In *Proceedings of ACL*.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Autoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*.

Jiajun Zhang and Chengqing Zong. 2013. Learning a phrase-based translation model from monolingual data with application to domain adaptation. In *Proceedings of ACL*.