Maximum Expected Likelihood Estimation for Zero-Resource Neural Machine Translation

Hao Zheng[#], **Yong Cheng**[+], **Yang Liu**[†‡*]
[#] Beihang University, Beijing, China
[+] 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
[‡]Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China
{mswellhao, chengyong3001}@gmail.com
liuyang2011@tsinghua.edu.cn

Abstract

While neural machine translation (NMT) has made remarkable progress in translating a handful of resource-rich language pairs recently, parallel corpora are not always readily available for most language pairs. To deal with this problem, we propose an approach to zero-resource NMT via *maximum expected likelihood estimation*. The basic idea is to maximize the expectation with respect to a pivot-to-source translation model for the intended source-to-target model on a pivot-target parallel corpus. To approximate the expectation, we propose two methods to connect the pivot-to-source and source-to-target models. Experiments on two zero-resource language pairs show that the proposed approach yields substantial gains over baseline methods. We also observe that when trained jointly with the source-to-target model, the pivot-to-source translation model also obtains improvements over independent training.

1 Introduction

Recently, neural machine translation (NMT) has achieved state-of-the-art performance on language pairs with abundant parallel corpora available [Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014]. Nevertheless, large-scale, high-quality parallel corpora are non-existent for most language pairs [Utiyama and Isahara, 2007]. Studies reveal that it is challenging for NMT to yield satisfactory results for resource-scarce language pairs [Zoph *et al.*, 2016; Firat *et al.*, 2016]. Zoph *et al.* [2016] indicate that NMT tends to learn degenerate estimates from low-count events, which deteriorates the translation performance of NMT under small-data conditions.

As a result, developing methods to achieve zero-resource neural machine translation, in which no direct source-target parallel corpora are available, has attracted increasing attention in the community recently [Johnson *et al.*, 2016; Firat *et al.*, 2016]. Existing work can be roughly divided into two broad categories: *multilingual* and *pivot-based* approaches. The multilingual approaches focus on leverging multilingual parallel corpora to achieve zero-resource NMT. Johnson *et al.* [2016] introduce a simple solution to use a single NMT model to translate between multiple languages. Requiring no change to the model architecture, their universal model takes advantage of multilingual data to improve NMT for all languages pairs involved. Firat *et al.* [2016] propose a finetuning algorithm for multiway, multilingual neural machine translation that enables zero-resource machine translation. Although these multilingual methods achieve direct source-to-target translation without parallel corpora available, it is fairly difficult to learn and analyze the universal representation for multiple languages.

Another important direction is to introduce a third language called *pivot* to bridge the source and target languages, which has been widely used in zero-resource statistical machine translation [Cohn and Lapata, 2007; Wu and Wang, 2007; Utiyama and Isahara, 2007; Bertoldi *et al.*, 2008; Habash and Hu, 2009]. Bertoldi *et al.* [2008] present an approach to constructing a pseudo source-target parallel corpus by translating the pivot sentences in the pivot-target corpus into the source language with the pivot-to-source model. On the other hand, one of the most representative pivot-based methods is the *pivot-based translation* [Utiyama and Isahara, 2007], which achieves source-to-target translation indirectly using the source-to-pivot and pivot-to-target models: a source sentence is first translated into a pivot sentence, which is then translated into a target sentence. Recently, Johnson *et al.* [2016] show the pivot-based translation for NMT significantly outperforms their universal NMT without incremental training. Although the pivot-based translation is a simple and effective approach to zero-resource NMT, they often suffer from the error propagation problem due to indirect modeling: mistakes made in source-to-pivot translation will be propagated to pivot-to-target translation.

In this work, we aim to achieve direct modeling of zero-resource source-to-target NMT and minimizing the requirement of multilingual data. We propose an approach to directly training the source-to-target translation model via maximum expected likelihood estimation. Our training objective is to maximize the expectation with respect to a pivot-to-source translation model for the intended source-to-target model on a pivot-target parallel corpus. The assumption underlying our idea is that if a pivot sentence $\mathbf{z}$ and a target sentence $\mathbf{y}$ constitute a parallel sentence pair, the source translation $\mathbf{x}$ of the pivot sentence should also be translationally equivalent to $\mathbf{y}$.

[*] Corresponding author: Yang Liu.

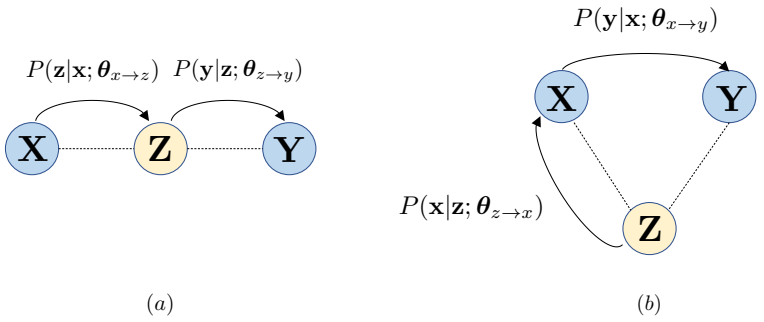

Figure 1: (a) The pivot-based approach and (b) the maximum expected likelihood estimation approach to zero-resource neural machine translation. $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ denote source, target, and pivot languages, respectively. We use a dished line to denote that there exists a parallel corpus available for the connected language pair. Solid lines with arrows represent translation directions. The pivot-based approach leverages a pivot to achieve indirect source-to-target translation: it first translates $\mathbf{x}$ to $\mathbf{z}$, which is then translated to $\mathbf{y}$. Maximum expected likelihood estimation aims to maximize the expectation with respect to a pivot-to-source translation model $P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_{z\to x})$ for the intended source-to-target model $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$ on a pivot-target parallel corpus.

Therefore, our approach involves two models: the *pivot-to-source* model that provides an artificial parallel corpus and the *source-to-target* model that is learned on the artificial data. As it is intractable to enumerate all sentences in the source language, we also propose two strategies to connect the pivot-to-source and source-to-target models. Experiments on two zero-resource translation tasks demonstrate that the proposed approach yields substantial gains over the baseline methods.

2 Background

2.1 Maximum Likelihood Estimation

Given a source-language sentence $\mathbf{x}$ and a target-language sentence $\mathbf{y}$, we use $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$ to denote a source-to-target NMT model [Bahdanau *et al.*, 2014].

For resource-rich language pairs, a source-target parallel corpus $D_{x,y} = \{\langle \mathbf{x}^{(k)}, \mathbf{y}^{(k)}\rangle\}_{k=1}^{K}$ is available to train the source-to-target NMT model. The standard training objective in this scenario is to maximize the log-likelihood of the training data:

$$\hat{\boldsymbol{\theta}}_{x\to y} = \underset{\boldsymbol{\theta}_{x\to y}}{\operatorname{argmax}} \Big\{ \mathcal{L}(\boldsymbol{\theta}_{x\to y}) \Big\} \tag{1}$$

where the log-likelihood is defined as

$$\mathcal{L}(\boldsymbol{\theta}_{x\to y}) = \sum_{k=1}^{K} \log P(\mathbf{y}^{(k)}|\mathbf{x}^{(k)};\boldsymbol{\theta}_{x\to y}) \tag{2}$$

2.2 Pivot-based Translation

For zero-resource language pairs, parallel corpora are usually not readily available. As a result, previous work has endeavored to leverage a third language called *pivot* to bridge the source and target languages. Let $\mathbf{z}$ be a pivot-language sentence. As shown in Figure 1, source-to-target translation can be modeled *indirectly* by cascading two sub-models:

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to z},\boldsymbol{\theta}_{z\to y}) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{x\to z}) P(\mathbf{y}|\mathbf{z};\boldsymbol{\theta}_{z\to y}) \tag{3}$$

The pivot-based translation assumes that a source-pivot parallel corpus $D_{x,z} = \{\langle \mathbf{x}^{(m)}, \mathbf{z}^{(m)}\rangle\}_{m=1}^{M}$ and a pivot-target parallel corpus $D_{z,y} = \{\langle \mathbf{z}^{(n)}, \mathbf{y}^{(n)}\rangle\}_{n=1}^{N}$ are available. As a result, the source-to-pivot and pivot-to-target models can be trained separately:

$$\hat{\boldsymbol{\theta}}_{x\to z} = \underset{\boldsymbol{\theta}_{x\to z}}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \log P(\mathbf{z}^{(m)}|\mathbf{x}^{(m)};\boldsymbol{\theta}_{x\to z}) \right\} \tag{4}$$

$$\hat{\boldsymbol{\theta}}_{z\to y} = \underset{\boldsymbol{\theta}_{z\to y}}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} \log P(\mathbf{y}^{(n)}|\mathbf{z}^{(n)};\boldsymbol{\theta}_{z\to y}) \right\} \tag{5}$$

Then, the two models can be used to perform source-to-target translation in two steps:

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \Big\{ P(\mathbf{z}|\mathbf{x};\hat{\boldsymbol{\theta}}_{x\to z}) \Big\} \tag{6}$$

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \Big\{ P(\mathbf{y}|\hat{\mathbf{z}};\hat{\boldsymbol{\theta}}_{z\to y}) \Big\} \tag{7}$$

One drawback of this approach is that the translation quality of source-to-target translation heavily depends on the selection of $\hat{\mathbf{z}}$, which often fails to retain exactly the same information with the original source sentence and thus leads to severe cascaded translation errors.

3 Approach

3.1 Maximum Expected Likelihood Estimation

As shown in Figure 1, our idea is to maximize the expectation with respect to a pivot-to-source translation model for the intended source-to-target model on a pivot-target parallel corpus:

$$\hat{\boldsymbol{\theta}}_{x\to y} = \underset{\boldsymbol{\theta}_{x\to y}}{\operatorname{argmax}} \Big\{ \mathcal{J}_{\text{indep}}(\boldsymbol{\theta}_{x\to y}) \Big\} \tag{8}$$

where the training objective is defined as

$$\mathcal{J}_{\text{indep}}(\boldsymbol{\theta}_{x\to y}) = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}|\mathbf{z}^{(n)};\hat{\boldsymbol{\theta}}_{z\to x}} \Big[\log P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y}) \Big] \tag{9}$$

Note that the pivot-to-source model is pre-trained on the source-pivot corpus:

$$\hat{\boldsymbol{\theta}}_{z\to x} = \underset{\boldsymbol{\theta}_{z\to x}}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\theta}_{z\to x}) \right\} \tag{10}$$

where the log-likelihood is defined as

$$\mathcal{L}(\boldsymbol{\theta}_{z\to x}) = \sum_{m=1}^{M} \log P(\mathbf{x}^{(m)}|\mathbf{z}^{(m)}; \boldsymbol{\theta}_{z\to x}) \tag{11}$$

We refer to Eq. (8) and (9) as **maximum expected likelihood estimation**.

Maximum expected likelihood estimation is an extension of standard maximum likelihood estimation for learning the *intended distribution* $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$ by introducing a *data generating distribution* $P(\mathbf{x}|\mathbf{z};\boldsymbol{\theta}_{z\to x})$. This new training criterion for zero-resource NMT is based on the following assumption: *since* $\langle \mathbf{z}^{(n)}, \mathbf{y}^{(n)} \rangle$ *is a bilingual sentence pair, the source translation of* $\mathbf{z}^{(n)}$ *(i.e.,* $\mathbf{x}$*) should also be translationally equivalent to* $\mathbf{y}^{(n)}$.

However, as calculating the expectation in Eq. (9) needs to enumerate all possible source sentences, it is intractable to optimize source-to-target model parameters due to the exponential search space. As a result, we propose two approximation methods to address this problem.

3.2 Approximation

Using Single Word Embeddings

Let $\mathcal{X}(\mathbf{z})$ be the set of all possible source translations of a pivot sentence $\mathbf{z}$. A standard solution is to approximate the full search space with a sampled subset:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}|\mathbf{z}^{(n)};\hat{\boldsymbol{\theta}}_{z\to x}}\Big[\log P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big] \\
&= \sum_{\mathbf{x}\in\mathcal{X}(\mathbf{z}^{(n)})} P(\mathbf{x}|\mathbf{z}^{(n)};\hat{\boldsymbol{\theta}}_{z\to x}) \log P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y}) \\
&\approx \sum_{\mathbf{x}\in\mathcal{S}(\mathbf{z}^{(n)})} P(\mathbf{x}|\mathbf{z}^{(n)};\hat{\boldsymbol{\theta}}_{z\to x}) \log P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y})
\end{aligned} \tag{12}$$

where $\mathcal{S}(\mathbf{z}) \subset \mathcal{X}(\mathbf{z})$ is a sampled subset.

Note that the intended distribution $P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$ actually takes the concatenation of word embeddings of $\mathbf{x}$ as input. Let $\mathbf{x} = x_1, \ldots, x_t, \ldots, x_T$ be a sampled source translation with T words for $\mathbf{z}^{(n)}$. We use $e(x_t) \in \mathbb{R}^{d\times 1}$ to represent the word embedding of x_t. Therefore, the vector representation of $\mathbf{x}$ is given by

$$e(\mathbf{x}) = \big\{e(x_t)\big\}_{t=1}^{T} \tag{13}$$

Therefore, Eq. (12) can be equivalently written as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}|\mathbf{z}^{(n)};\hat{\boldsymbol{\theta}}_{z\to x}}\Big[\log P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big] \\
&\approx \sum_{\mathbf{x}\in\mathcal{S}(\mathbf{z}^{(n)})} P\big(e(\mathbf{x})|e(\mathbf{z}^{(n)});\hat{\boldsymbol{\theta}}_{z\to x}\big) \times \\
&\qquad \log P\big(e(\mathbf{y}^{(n)})|e(\mathbf{x});\boldsymbol{\theta}_{x\to y}\big)\big)
\end{aligned} \tag{14}$$

Although approximation with sampling proves to achieve a reasonable balance between effectiveness and efficiency [Shen *et al.*, 2016], the sampled space is still restricted to a limited number of candidate translations. In addition, only a small fraction of words in the entire vocabulary are included in the samples. As a result, our approach also potentially faces the error propagation problem: sampling mistakes will affect the estimation of intended model parameters.

Using Expected Word Embeddings

Inspired by [Kočiský *et al.*, 2016], we propose to use *expected word embeddings* rather than single word embeddings to circumvent this drawback. Given a sampled source translation $\mathbf{x}^{(s)} \in \mathcal{S}(\mathbf{z}^{(n)})$, at each time step in the decoder of the pivot-to-source model, an expected word embedding for the t-th source word x_t is calculated as

$$\begin{aligned}
&\mathbb{E}_{x|\mathbf{z}^{(n)},\mathbf{x}^{(s)}_{<t};\hat{\boldsymbol{\theta}}_{z\to x}}\big[e(x)\big] \\
&= \sum_{x\in\mathcal{V}_x} P(x|\mathbf{z}^{(n)},\mathbf{x}^{(s)}_{<t};\hat{\boldsymbol{\theta}}_{z\to x}) e(x)
\end{aligned} \tag{15}$$

where $\mathcal{V}_x$ is the vocabulary of the source language.

As a result, provided with a sampled source sentence $\mathbf{x}^{(s)}$, the expected vector representation of a source sentence $\mathbf{x}$ can be approximated with the concatenation of expected word embeddings, which is defined as

$$\begin{aligned}
&\mathcal{E}(\mathbf{x}^{(s)}, \mathbf{z}^{(n)}, \mathcal{V}_x, \hat{\theta}_{z\to x}) \\
&= \Big\{\mathbb{E}_{x|\mathbf{z}^{(n)},\mathbf{x}^{(s)}_{<t};\hat{\boldsymbol{\theta}}_{z\to x}}\big[e(x)\big]\Big\}_{t=1}^{T}
\end{aligned} \tag{16}$$

Note that $\mathcal{E}(\mathbf{x}^{(s)}, \mathbf{z}^{(n)}, \mathcal{V}_x, \hat{\theta}_{z\to x})$ depends on the selection of $\mathbf{x}^{(s)}$.

As the expected word embeddings consider the entire vocabulary, we can leverage the expected word embeddings to implicitly represent the full search space $\mathcal{X}(\mathbf{z}^{(n)})$ approximately:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}|\mathbf{z}^{(n)};\hat{\boldsymbol{\theta}}_{z\to x}}\Big[\log P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big] \\
&\approx \frac{1}{|\mathcal{S}(\mathbf{z}^{(n)})|} \times \\
&\sum_{\mathbf{x}^{(s)}\in\mathcal{S}(\mathbf{z}^{(n)})} \log P\Big(e(\mathbf{y}^{(n)})|\mathcal{E}(\mathbf{x}^{(s)}, \mathbf{z}^{(n)}, \mathcal{V}_x, \hat{\theta}_{z\to x});\boldsymbol{\theta}_{x\to y}\Big)
\end{aligned}$$

Besides taking the full vocabulary into consideration, another important advantage of using expected word embeddings over the single embeddings counterpart is that pivot-to-source and source-to-target models are connected more closely. Consider Eq. (14). It is clear that the partial derivatives with respect to the parameters of the two models are independent of each other. In contrast, using expected word embeddings explicitly makes the calculation of partial derivatives of one model dependent on another model since it allows the error from the source-to-target model to be back-propagated to the pivot-to-source model through the expected word embeddings.

Therefore, we propose to use joint training to enable the interaction between pivot-to-source and source-to-target models:

$$\hat{\boldsymbol{\theta}}_{z\to x}, \hat{\boldsymbol{\theta}}_{x\to y} = \underset{\boldsymbol{\theta}_{z\to x},\boldsymbol{\theta}_{x\to y}}{\operatorname{argmax}} \left\{ \mathcal{J}_{\text{joint}}(\boldsymbol{\theta}_{z\to x}, \boldsymbol{\theta}_{x\to y}) \right\} \tag{17}$$

	Spanish-English		German-English		English-French	
	Es	En	Ge	En	En	Fr
# Sent.	850K		840K		900K	
# Word	23.23M	21.44M	20.88M	21.91M	22.56M	25.00M

Table 1: Statistics of parallel corpora used in our experiments. We evaluate our approach on two zero-resource translation tasks: Spanish-French and German-French. English is used as the pivot language.

where the new training objective is given by

$$\begin{aligned} & \mathcal{J}_{\text{joint}}(\boldsymbol{\theta}_{z\to x}, \boldsymbol{\theta}_{x\to y}) \\ = & \sum_{n=1}^{N} \mathbb{E}_{\mathbf{x}|\mathbf{z}^{(n)};\boldsymbol{\theta}_{z\to x}} \Big[\log P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y}) \Big] \\ & + \mathcal{L}(\boldsymbol{\theta}_{z\to x}) \end{aligned} \tag{18}$$

4 Experiments

4.1 Setup

Data Preparation

We evaluate our approach on two zero-resource translation tasks:

1. Spanish-English-French: Spanish as the source language, English as the pivot language, and French as the target language;
2. German-English-French: German as the source language, English as the pivot language, and French as the target language.

We use Spanish-English, German-English and English-French parallel corpora from the Europarl dataset. For each language pair, we retain sentence pairs with no more than 50 words. We also split the overlapping part of pivot sentences in the source-pivot and pivot-target corpora into two separate parts with equal size, which are then merged with the non-overlapping parts of the source-pivot and pivot-target corpora, respectively. Removing overlapped data ensures that none of the source-target parallel sentence pairs is present in our training data. All sentences are tokenized by the *tokenize.perl* script [Koehn *et al.*, 2007].

Table 1 shows the detailed statistics of parallel corpora used in our experiments. For Spanish-English, the training data contains 850K sentence pairs with 23.23M Spanish words and 21.44M English words. For German-English, the training data consists of 840K sentence pairs with 20.88M German words and 21.91M English words. They share the same English-French corpus which has 900K sentence pairs with 22.56M English words and 25.00M French words. The shared task 2006 datasets are used as development and test sets. The evaluation metric is case-insensitive BLEU [Papineni *et al.*, 2002], calculated by the *multi-bleu.perl* script.

Baseline Methods

We compare our approach with the following baseline methods:

1. PSEUDO [Bertoldi *et al.*, 2008]: a method for building pseudo parallel corpora adapted for NMT.
2. PIVOT [Utiyama and Isahara, 2007]: a pivot-based translation method adapted for NMT.

For the PSEUDO method, we first train a pivot-to-source NMT model $P(\mathbf{x}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to x})$ on the source-pivot parallel corpus. Then, the pivot sentences $\{\mathbf{z}^{(n)}\}_{n=1}^{N}$ in the pivot-target parallel corpus $D_{z,y} = \{\langle \mathbf{z}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$ are translated into source sentences $\{\tilde{\mathbf{x}}^{(n)}\}_{n=1}^{N}$ using the pivot-to-source NMT model $P(\mathbf{x}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to x})$. These source sentences $\{\tilde{\mathbf{x}}^{(n)}\}_{n=1}^{N}$ and the original target sentences $\{\mathbf{y}^{(n)}\}_{n=1}^{N}$ in the pivot-target parallel corpus constitute a pseudo source-target parallel corpus $D_{\tilde{x},y} = \{\langle \tilde{\mathbf{x}}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$, which can be used to train the source-to-target model $P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{x\to y})$. The PSEUDO method can be treated as a special case of maximum expected likelihood estimation since only the candidate translation with the highest probability in the full space is considered.

For the PIVOT method, we first train source-to-pivot and pivot-to-target NMT models separately on source-pivot and pivot-target parallel corpora. In decoding, a source sentence is first translated into a pivot sentence using the source-to-pivot model, which is then translated to a target sentence using the pivot-to-target model. Please refer to Section 2.2 for more details. This approach serves as the baseline in previous work on zero-source neural machine translation task [Johnson *et al.*, 2016; Firat *et al.*, 2016]. It is worth emphasizing that the PIVOT method is a very strong baseline as Johnson *et al.* [2016] show that it yields much higher BLEU scores than their universal NMT model without incremental training and the direct source-to-target translation model proposed in [Firat *et al.*, 2016] alone also performs worse than the PIVOT method.

We implement all methods on top of the state-of-the-art open-source NMT system GROUNDHOG [Bahdanau *et al.*, 2014]. All neural translation models use the default setting of network hyper-parameters of GROUNDHOG.

Training Details

Our joint training approach requires the parameters of multiple translation models to be updated jointly. In practice, we find that it is extremely slow for the models to converge from random initialization. Inspired by previous work on training multi-layered perceptrons [Bengio *et al.*, 2007], we use the weights of the pivot-to-source model trained independently and the source-to-target model trained with the PSEUDO method to initialize joint training. For fair comparison, we also initialize the source-to-target model in our independent training approach with the model obtained by the PSEUDO method. Although each mini-batch takes more time to update than standard MLE training, our approaches converge very fast thanks to the aforementioned initialization

Method	Embed.	Training	Spanish-French Dev.	Spanish-French Test	German-French Dev.	German-French Test
PSEUDO	single	indep.	28.04	28.27	20.23	19.92
PIVOT	single	indep.	29.51	29.86	23.49	23.33
this work	single	indep.	32.48	32.27	24.37	24.18
	expected	indep.	32.98	33.04	24.41	24.49
		joint	33.97	33.83	25.41	24.99

Table 2: Comparison with PSEUDO and PIVOT on the Europarl dataset. We evaluate our approach on two zero-resource translation tasks: Spanish-French and German-French. English is used as the pivot language. PSEUDO denotes using the pivot-to-source NMT model trained on the source-pivot parallel corpus to translate the pivot sentences into source sentences, which is combined with the target part in the pivot-target corpus to form a pseudo source-target parallel corpus [Bertoldi *et al.*, 2008]. PIVOT denotes using the source-to-pivot model to translate a source sentence into a pivot sentence, which is then translated into a target sentence with the pivot-to-target model [Utiyama and Isahara, 2007]. "Embed." indicates the way to form the vector representation of a sentence using word embeddings (see Section 3.2 for details). "Training" indicates whether two sub-models are trained jointly or not. The evaluation metric is case-insensitive BLEU.

techniques.

Each time the sentence $\mathbf{z}^{(n)}$ is selected in a mini-batch, we only randomly sample one sentence $\mathbf{x}$ in consideration of the limited GPU memory. Note that a set of different sentences will still be sampled for each $\mathbf{z}^{(n)}$ as $\mathbf{z}^{(n)}$ is usually selected in mini-batches multiple times. This is important since it can effectively lower the variance of approximation. For the single embedding approach, we find the probability weight $P(\mathbf{x}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to x})$ is usually very small, making the training extremely slow. Therefore, in practice we instead take its q-th root $\sqrt[q]{P(\mathbf{x}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\to x})}$ and set $q = 10$ for speed-up.

4.2 Comparison with PSEUDO and PIVOT

Table 2 shows the comparison of our approach with PSEUDO and PIVOT on the Europarl corpus. PSEUDO achieves a BLEU score of 28.27 on the Spanish-French task and 19.92 on German-French, which are much lower than those of other approaches. This is because PSEUDO only uses the source translation with the highest probability to build the pseudo parallel corpus, which may also cause severe error propagation problem in training: mistakes made in the pivot-to-source translation affect the quality of pseudo parallel corpus.

The PIVOT approach significantly outperforms PSEUDO for both translation tasks. Although PIVOT also faces the error propagation problem in decoding, the source-pivot and pivot-target parallel corpora used to train source-to-pivot and pivot-to-target NMT models are supposed to be clean. In contrast, the source part of the pseudo parallel corpus inevitably contains much noise due to translation errors.

Our approach significantly outperforms both PSEUDO and PIVOT on both tasks. Note that PSEUDO can be considered as a special case of maximum expected likelihood estimation: only the candidate with the highest probability in the full space is used. This finding suggests that it is important to take multiple candidates into consideration in maximum expected likelihood estimation. As using expected embeddings is capable of exploiting the full vocabulary, it outperforms using single word embeddings significantly.

Another important finding is that joint training leads to significant improvements over independent training thanks to the interaction between the pivot-to-source and source-to-target models during training. Table 3 shows that the interaction not only improves source-to-target translation, but also enhances pivot-to-source translation. The BLEU scores increase by over 2% for both tasks.

Training	English-Spanish Dev.	English-Spanish Test	English-German Dev.	English-German Test
Indep.	30.73	31.16	19.61	19.56
Joint	31.83	32.57	21.99	21.77

Table 3: Comparison between independent and joint training on pivot-to-source translation.

Direction	Spanish-French Dev.	Spanish-French Test	German-French Dev.	German-French Test
unidirectional	33.97	33.83	25.41	24.99
bidirectional	34.66	34.41	25.62	25.42

Table 4: Effect of bidirectional training.

4.3 Effect of Bidirectional Training

From the perspective of pivot-based approaches, our approach can be seen as indirectly modeling pivot-to-target translation via pivot-to-source and source-to-target translation models on the pivot-target parallel corpus $D_{z,y} = \{\langle \mathbf{z}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{m=1}^{M}$.

$$\begin{aligned} & P(\mathbf{y}^{(n)}|\mathbf{z}^{(n)};\boldsymbol{\theta}_{z\to x},\boldsymbol{\theta}_{x\to y}) \\ = & \sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{z}^{(n)};\boldsymbol{\theta}_{z\to x}) P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y}) \\ = & \mathbb{E}_{\mathbf{x}|\mathbf{z}^{(n)};\boldsymbol{\theta}_{z\to x}}\Big[P(\mathbf{y}^{(n)}|\mathbf{x};\boldsymbol{\theta}_{x\to y})\Big] \end{aligned} \tag{19}$$

An alternative is to make use of the source-pivot parallel

Method	Training Data					Spanish-French		German-French	
	es-en	de-en	en-fr	es-fr	de-fr	Dev.	Test	Dev.	Test
ORACLE	-	-	-	100K	100K	24.26	24.62	13.58	13.68
	-	-	-	1.0M	1.0M	33.99	33.33	23.83	23.41
	-	-	-	1.5M	1.5M	36.81	36.22	25.78	25.54
this work	850K	840K	900K	-	-	34.66	34.41	25.62	25.42

Table 5: Comparison with ORACLE that uses direct source-target parallel corpora.

corpus $D_{x,z} = \{\langle \mathbf{x}^{(n)}, \mathbf{z}^{(n)} \rangle\}_{n=1}^{N}$:

$$\begin{aligned} & P(\mathbf{z}^{(m)}|\mathbf{x}^{(m)}; \boldsymbol{\theta}_{x\to y}, \boldsymbol{\theta}_{y\to z}) \\ = & \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}^{(m)}; \boldsymbol{\theta}_{x\to y}) P(\mathbf{z}^{(m)}|\mathbf{y}; \boldsymbol{\theta}_{y\to z}) \\ = & \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(m)};\boldsymbol{\theta}_{x\to y}}\Big[P(\mathbf{z}^{(m)}|\mathbf{y}; \boldsymbol{\theta}_{y\to z})\Big] \end{aligned} \quad (20)$$

For convenience, we refer to Eq. (19) as the $z \to x \to y$ direction and Eq. (20) as the $x \to y \to z$ direction.

The joint training objective for the $x \to y \to z$ direction is given by

$$\begin{aligned} & \mathcal{J}_{\text{joint}}(\boldsymbol{\theta}_{x\to y}, \boldsymbol{\theta}_{y\to z}) \\ = & \sum_{m=1}^{M} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(m)};\boldsymbol{\theta}_{x\to y}}\Big[\log P(\mathbf{z}^{(m)}|\mathbf{y}; \boldsymbol{\theta}_{y\to z})\Big] + \\ & \mathcal{L}(\boldsymbol{\theta}_{y\to z}) \end{aligned} \quad (21)$$

As bidirectional information has proven to be complementary for the learning of mapping between two languages [Och and Ney, 2002; Li and Jurafsky, 2016], the two directions can be combined to perform *bidirectional training*:

$$\begin{aligned} & \mathcal{B}(\boldsymbol{\theta}_{x\to y}, \boldsymbol{\theta}_{y\to z}, \boldsymbol{\theta}_{z\to x}) \\ = & \mathcal{J}_{\text{joint}}(\boldsymbol{\theta}_{z\to x}, \boldsymbol{\theta}_{x\to y}) + \lambda \mathcal{J}_{\text{joint}}(\boldsymbol{\theta}_{x\to y}, \boldsymbol{\theta}_{y\to z}) \end{aligned} \quad (22)$$

where λ is a hyper-parameter that balance the preference between two directions. We set $\lambda = 0.1$ in our experiments.

Table 4 shows the effect of bidirectional training. "unidirectional" denotes the $z \to x \to y$ direction and "bidirectional" denotes combining the the $z \to x \to y$ and $x \to y \to z$ directions. Combining the two directions leads to significant improvements ($p < 0.01$). Note that training alone in the $x \to y \to z$ direction results in very low BLEU scores. One possible reason is that the training objective $\mathcal{J}_{\text{joint}}(\boldsymbol{\theta}_{x\to y}, \boldsymbol{\theta}_{y\to z})$ serves as more like a regularization item, in the sense that it can provide some useful information to further improve the model's performance, but it alone lacks the capability to provide complete and effective supervision.

4.4 Comparison with ORACLE

Table 5 compares our best approach with ORACLE that uses direct source-target parallel corpora. We observe that our approach outperforms ORACLE with 100K and 1M parallel sentences. This is very encouraging since our approach only uses three source-pivot and pivot-target corpora with around 850K parallel sentences.

5 Related Work

Recently, zero-resource neural machine translation has received much attention in the community. Firat *et al.* [2016] pre-train multi-way multilingual model and then fine-tune it with pseudo parallel data generated by the model. However, the use of pseudo parallel data as supervision unavoidably suffers from the error propagation problem. This is probably why their one-to-one direct translation model performs worse than the pivot-based translation strategy. Google's multilingual neural machine translation system uses a single NMT model to translate between multiple languages, naturally enabling zero-resource translation [Johnson *et al.*, 2016].

A wide variety of approaches have been proposed for zero-resource or low-resource translation in conventional SMT systems. Wu et al. [2007] and Cohn et al. [2007] use source-to-pivot and pivot-to-target translation models to induce a new source-to-target phrase table, on which a source-to-target translation model is built. Bertoldi et al. [2008] exploit existing models to build a pseudo source-target parallel corpus, from which a source-to-target model can be trained. These methods fall into the broad category of directly constructing a source-to-target translation model. Utiyama and Isahara [2007] compare two translation strategies, namely phrase translation and sentence translation. The former is similar to the work in [Wu and Wang, 2007], while the latter is the pivot-based translation which first translates the source sentence to the pivot sentence and then to the target sentence. Cheng *et al.* [2016] improve the pivot-based translation by jointly training the source-to-pivot and pivot-to-target translation models. Nakayama and Nishida [2016] perform zero-resource machine translation using multimedia as the pivot.

Our work is in spirit close to Bertoldi et al. [2008] since both assume that if a pivot sentence $\mathbf{z}$ and a target sentence $\mathbf{y}$ constitute a parallel sentence pair, the source translation $\mathbf{x}$ of the pivot sentence $\mathbf{z}$ should also be translationally equivalent to $\mathbf{y}$. However, we formally define the expected training objective and adapt it to NMT, thus enabling the novel expected embedding and joint training strategies.

6 Conclusion

We have presented a method for training neural machine translation models for zero-resource language pairs using maximum expected likelihood estimation. The central idea is to leverage a data generating distribution to provide pseudo parallel corpora for training the intended distribution. Experiments on the Europarl corpus show that our approach achieves significant improvements over the pivot-based translation approach.

Acknowledgments

This work was done while Hao Zheng was visiting Tsinghua University. This work is supported by the National Natural Science Foundation of China (No.61522204), the 863 Program (2015AA015407), and the National Natural Science Foundation of China (No.61432013). This research is also supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme. We thank the anonymous reviewers for their insightful comments.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2014.

[Bengio *et al.*, 2007] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Proceedings of NIPS*. 2007.

[Bertoldi *et al.*, 2008] Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. Phrase-based statistical machine translation with pivot languages. In *IWSLT*, 2008.

[Cheng *et al.*, 2016] Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. Neural machine translation with pivot languages. *arXiv preprint arXiv:1611.04928*, 2016.

[Cohn and Lapata, 2007] Trevor Cohn and Mirella Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of ACL*, 2007.

[Firat *et al.*, 2016] Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of EMNLP*, 2016.

[Habash and Hu, 2009] Nizar Habash and Jun Hu. Improving arabic-chinese statistical machine translation using english as pivot language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009.

[Johnson *et al.*, 2016] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

[Koehn *et al.*, 2007] 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. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, 2007.

[Kočiský *et al.*, 2016] Tomáš Kočiský, Gábor Melis, Edward Grefenstette, Chris Dyer, Wang Ling, Phil Blunsom, and Karl Moritz Hermann. Semantic parsing with semi-supervised sequential autoencoders. In *Proceedings of EMNLP*, 2016.

[Li and Jurafsky, 2016] Jiwei Li and Dan Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016.

[Nakayama and Nishida, 2016] Hideki Nakayama and Noriki Nishida. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *arXiv preprint arXiv:1611.04503*, 2016.

[Och and Ney, 2002] Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, 2002.

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

[Shen *et al.*, 2016] Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. 2016.

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

[Utiyama and Isahara, 2007] Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of HLT-NAACL*, 2007.

[Wu and Wang, 2007] Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 2007.

[Zoph *et al.*, 2016] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP*, 2016.