A Teacher-Student Framework for
Zero-Resource Neural Machine Translation

Yun Chen[†]**, Yang Liu**[‡*]**, Yong Cheng**[+]**, Victor O.K. Li**[†]

[†]Department of Electrical and Electronic Engineering, The University of Hong Kong
[‡]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
[+]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
`yun.chencreek@gmail.com; liuyang2011@tsinghua.edu.cn;`
`chengyong3001@gmail.com; vli@eee.hku.hk`

Abstract

While end-to-end neural machine translation (NMT) has made remarkable progress recently, it still suffers from the data scarcity problem for low-resource language pairs and domains. In this paper, we propose a method for zero-resource NMT by assuming that parallel sentences have close probabilities of generating a sentence in a third language. Based on the assumption, our method is able to train a source-to-target NMT model ("student") without parallel corpora available guided by an existing pivot-to-target NMT model ("teacher") on a source-pivot parallel corpus. Experimental results show that the proposed method significantly improves over a baseline pivot-based model by +3.0 BLEU points across various language pairs.

1 Introduction

Neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015), which directly models the translation process in an end-to-end way, has attracted intensive attention from the community. Although NMT has achieved state-of-the-art translation performance on resource-rich language pairs such as English-French and German-English (Luong et al., 2015; Jean et al., 2015; Wu et al., 2016; Johnson et al., 2016), it still suffers from the unavailability of large-scale parallel corpora for translating low-resource languages. Due to the large parameter space, neural models usually learn poorly from low-count events, resulting in a poor choice for low-resource language pairs. Zoph et

al. (2016) indicate that NMT obtains much worse translation quality than a statistical machine translation (SMT) system on low-resource languages.

As a result, a number of authors have endeavored to explore methods for translating language pairs without parallel corpora available. These methods can be roughly divided into two broad categories: *multilingual* and *pivot-based*. Firat et al. (2016b) present a multi-way, multilingual model with shared attention to achieve zero-resource translation. They fine-tune the attention part using pseudo bilingual sentences for the zero-resource language pair. Another direction is to develop a universal NMT model in multilingual scenarios (Johnson et al., 2016; Ha et al., 2016). They use parallel corpora of multiple languages to train one single model, which is then able to translate a language pair without parallel corpora available. Although these approaches prove to be effective, the combination of multiple languages in modeling and training leads to increased complexity compared with standard NMT.

Another direction is to achieve source-to-target NMT without parallel data via a *pivot*, which is either text (Cheng et al., 2016a) or image (Nakayama and Nishida, 2016). Cheng et al. (2016a) propose a pivot-based method for zero-resource NMT: it first translates the source language to a pivot language, which is then translated to the target language. Nakayama and Nishida (2016) show that using multimedia information as pivot also benefits zero-resource translation. However, pivot-based approaches usually need to divide the decoding process into two steps, which is not only more computationally expensive, but also potentially suffers from the error propagation problem (Zhu et al., 2013).

In this paper, we propose a new method for zero-resource neural machine translation. Our
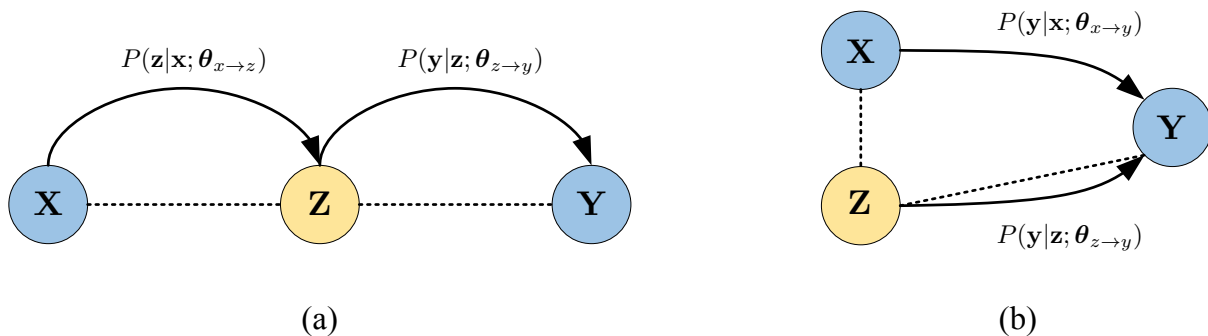
---

*Corresponding author: Yang Liu.

Figure 1: (a) The pivot-based approach and (b) the teacher-student approach to zero-resource neural machine translation. $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ denote source, target, and pivot languages, respectively. We use a dashed line to denote that there is 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 $\mathbf{x}$ into $\mathbf{z}$, which is then translated into $\mathbf{y}$. Our training algorithm is based on the translation equivalence assumption: if $\mathbf{x}$ is a translation of $\mathbf{z}$, then $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y})$ should be close to $P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z \to y})$. Our approach directly trains the intended source-to-target model $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y})$ ("student") on a source-pivot parallel corpus, with the guidance of an existing pivot-to-target model $P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z \to y})$ ("teacher").

method assumes that parallel sentences should have close probabilities of generating a sentence in a third language. To train a source-to-target NMT model without parallel corpora ("student"), we leverage an existing pivot-to-target NMT model ("teacher") to guide the learning process of the student model on a source-pivot parallel corpus. Compared with pivot-based approaches (Cheng et al., 2016a), our method allows direct parameter estimation of the intended NMT model, without the need to divide decoding into two steps. This strategy not only improves efficiency but also avoids error propagation in decoding. Experiments on the Europarl and WMT datasets show that our approach achieves significant improvements in terms of both translation quality and decoding efficiency over a baseline pivot-based approach to zero-resource NMT on Spanish-French and German-French translation tasks.

2 Background

Neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015) advocates the use of neural networks to model the translation process in an end-to-end manner. As a data-driven approach, NMT treats parallel corpora as the major source for acquiring translation knowledge.

Let $\mathbf{x}$ be a source-language sentence and $\mathbf{y}$ be a target-language sentence. We use $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y})$

to denote a source-to-target neural translation model, where $\boldsymbol{\theta}_{x \to y}$ is a set of model parameters. Given a source-target parallel corpus $D_{x,y}$, which is a set of parallel source-target sentences, the model parameters can be learned by maximizing the log-likelihood of the parallel corpus:

$$\hat{\boldsymbol{\theta}}_{x \to y} = \underset{\boldsymbol{\theta}_{x \to y}}{\arg\max} \left\{ \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in D_{x,y}} \log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y}) \right\}.$$

Given learned model parameters $\hat{\boldsymbol{\theta}}_{x \to y}$, the decision rule for finding the translation with the highest probability for a source sentence $\mathbf{x}$ is given by

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\arg\max} \left\{ P(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{x \to y}) \right\}. \quad (1)$$

As a data-driven approach, NMT heavily relies on the availability of large-scale parallel corpora to deliver state-of-the-art translation performance (Wu et al., 2016; Johnson et al., 2016). Zoph et al. (2016) report that NMT obtains much lower BLEU scores than SMT if only small-scale parallel corpora are available. Therefore, the heavy dependence on the quantity of training data poses a severe challenge for NMT to translate zero-resource language pairs.

Simple and easy-to-implement, pivot-based methods have been widely used in SMT for

translating zero-resource language pairs (de Gispert and Mariño, 2006; Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Wu and Wang, 2007; Bertoldi et al., 2008; Wu and Wang, 2009; Zahabi et al., 2013; Kholy et al., 2013). As pivot-based methods are agnostic to model structures, they have been adapted to NMT recently (Cheng et al., 2016a; Johnson et al., 2016).

Figure 1(a) illustrates the basic idea of pivot-based approaches to zero-resource NMT (Cheng et al., 2016a). Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ denote source, target, and pivot languages. We use dashed lines to denote language pairs with parallel corpora available and solid lines with arrows to denote translation directions.

Intuitively, the source-to-target translation can be indirectly modeled by bridging two NMT models via a pivot:

$$P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to z}, \boldsymbol{\theta}_{z \to y})$$
$$= \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x \to z}) P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z \to y}). \quad (2)$$

As shown in Figure 1(a), pivot-based approaches assume that the source-pivot parallel corpus $D_{x,z}$ and the pivot-target parallel corpus $D_{z,y}$ are available. As it is impractical to enumerate all possible pivot sentences, the two NMT models are trained separately in practice:

$$\hat{\boldsymbol{\theta}}_{x \to z} = \underset{\theta_{x \to z}}{\operatorname{argmax}} \left\{ \sum_{\langle \mathbf{x}, \mathbf{z} \rangle \in D_{x,z}} \log P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{x \to z}) \right\},$$

$$\hat{\boldsymbol{\theta}}_{z \to y} = \underset{\theta_{z \to y}}{\operatorname{argmax}} \left\{ \sum_{\langle \mathbf{z}, \mathbf{y} \rangle \in D_{z,y}} \log P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z \to y}) \right\}.$$

Due to the exponential search space of pivot sentences, the decoding process of translating an unseen source sentence $\mathbf{x}$ has to be divided into two steps:

$$\hat{\mathbf{z}} = \underset{\mathbf{z}}{\operatorname{argmax}} \left\{ P(\mathbf{z}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{x \to z}) \right\}, \quad (3)$$

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \left\{ P(\mathbf{y}|\hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}_{z \to y}) \right\}. \quad (4)$$

The above two-step decoding process potentially suffers from the error propagation problem (Zhu et al., 2013): the translation errors made in the first step (i.e., source-to-pivot translation) will affect the second step (i.e., pivot-to-target translation).

Therefore, it is necessary to explore methods to directly model source-to-target translation without parallel corpora available.

3 Approach

3.1 Assumptions

In this work, we propose to directly model the intended source-to-target neural translation based on a teacher-student framework. The basic idea is to use a pre-trained pivot-to-target model ("teacher") to guide the learning process of a source-to-target model ("student") without training data available on a source-pivot parallel corpus. One advantage of our approach is that Equation (1) can be used as the decision rule for decoding, which avoids the error propagation problem faced by two-step decoding in pivot-based approaches.

As shown in Figure 1(b), we still assume that a source-pivot parallel corpus $D_{x,z}$ and a pivot-target parallel corpus $D_{z,y}$ are available. Unlike pivot-based approaches, we first use the pivot-target parallel corpus $D_{z,y}$ to obtain a **teacher model** $P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z \to y})$, where $\hat{\boldsymbol{\theta}}_{z \to y}$ is a set of learned model parameters. Then, the teacher model "teaches" the **student model** $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y})$ on the source-pivot parallel corpus $D_{x,z}$ based on the following assumptions.

Assumption 1 *If a source sentence $\mathbf{x}$ is a translation of a pivot sentence $\mathbf{z}$, then the probability of generating a target sentence $\mathbf{y}$ from $\mathbf{x}$ should be close to that from its counterpart $\mathbf{z}$.*

We can further introduce a word-level assumption:

Assumption 2 *If a source sentence $\mathbf{x}$ is a translation of a pivot sentence $\mathbf{z}$, then the probability of generating a target word y from $\mathbf{x}$ should be close to that from its counterpart $\mathbf{z}$, given the already obtained partial translation $\mathbf{y}_{<j}$.*

The two assumptions are empirically verified in our experiments (see Table 2). In the following subsections, we will introduce two approaches to zero-resource neural machine translation based on the two assumptions.

3.2 Sentence-Level Teaching

Given a source-pivot parallel corpus $D_{x,z}$, our training objective based on Assumption 1 is defined as follows:

$$\mathcal{J}_{\text{SENT}}(\boldsymbol{\theta}_{x \to y})$$
$$= \sum_{\langle \mathbf{x}, \mathbf{z} \rangle \in D_{x,z}} \text{KL}\Big(P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z \to y}) \Big\| P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y}) \Big), \quad (5)$$

where the KL divergence sums over all possible target sentences:

$$\mathrm{KL}\Big(P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z\rightarrow y}) \big\| P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow y}) \Big)$$
$$= \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z\rightarrow y}) \log \frac{P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z\rightarrow y})}{P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow y})}. (6)$$

As the teacher model parameters are fixed, the training objective can be equivalently written as

$$\mathcal{J}_{\mathrm{SENT}}(\boldsymbol{\theta}_{x\rightarrow y})$$
$$= -\sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\rightarrow y}} \Big[\log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x\rightarrow y}) \Big]. \quad (7)$$

In training, our goal is to find a set of source-to-target model parameters that minimizes the training objective:

$$\hat{\boldsymbol{\theta}}_{x\rightarrow y} = \underset{\boldsymbol{\theta}_{x\rightarrow y}}{\operatorname{argmin}} \Big\{ \mathcal{J}_{\mathrm{SENT}}(\boldsymbol{\theta}_{x\rightarrow y}) \Big\}. \quad (8)$$

With learned source-to-target model parameters $\hat{\boldsymbol{\theta}}_{x\rightarrow y}$, we use the standard decision rule as shown in Equation (1) to find the translation $\hat{\mathbf{y}}$ for a source sentence $\mathbf{x}$.

However, a major difficulty faced by our approach is the intractability in calculating the gradients because of the exponential search space of target sentences. To address this problem, it is possible to construct a sub-space by either sampling (Shen et al., 2016), generating a k-best list (Cheng et al., 2016b) or mode approximation (Kim and Rush, 2016). Then, standard stochastic gradient descent algorithms can be used to optimize model parameters.

3.3 Word-Level Teaching

Instead of minimizing the KL divergence between the teacher and student models at the sentence level, we further define a training objective at the word level based on Assumption 2:

$$\mathcal{J}_{\mathrm{WORD}}(\boldsymbol{\theta}_{x\rightarrow y})$$
$$= \sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\rightarrow y}} \Big[J(\mathbf{x}, \mathbf{y}, \mathbf{z}, \hat{\boldsymbol{\theta}}_{z\rightarrow y}, \boldsymbol{\theta}_{x\rightarrow y}) \Big], \quad (9)$$

where

$$J(\mathbf{x}, \mathbf{y}, \mathbf{z}, \hat{\boldsymbol{\theta}}_{z\rightarrow y}, \boldsymbol{\theta}_{x\rightarrow y})$$
$$= \sum_{j=1}^{|\mathbf{y}|} \mathrm{KL}\Big(P(y|\mathbf{z}, \mathbf{y}_{<j}; \hat{\boldsymbol{\theta}}_{z\rightarrow y}) \big\|$$
$$P(y|\mathbf{x}, \mathbf{y}_{<j}; \boldsymbol{\theta}_{x\rightarrow y}) \Big). \quad (10)$$

Equation (9) suggests that the teacher model $P(y|\mathbf{z}, \mathbf{y}_{<j}; \hat{\boldsymbol{\theta}}_{z\rightarrow y})$ "teaches" the student model $P(y|\mathbf{x}, \mathbf{y}_{<j}; \boldsymbol{\theta}_{x\rightarrow y})$ in a word-by-word way. Note that the KL-divergence between two models is defined at the word level:

$$\mathrm{KL}\Big(P(y|\mathbf{z}, \mathbf{y}_{<j}; \hat{\boldsymbol{\theta}}_{z\rightarrow y}) \big\| P(y|\mathbf{x}, \mathbf{y}_{<j}; \boldsymbol{\theta}_{x\rightarrow y}) \Big)$$
$$= \sum_{y \in \mathcal{V}_y} P(y|\mathbf{z}, \mathbf{y}_{<j}; \hat{\boldsymbol{\theta}}_{z\rightarrow y}) \log \frac{P(y|\mathbf{z}, \mathbf{y}_{<j}; \hat{\boldsymbol{\theta}}_{z\rightarrow y})}{P(y|\mathbf{x}, \mathbf{y}_{<j}; \boldsymbol{\theta}_{x\rightarrow y})},$$

where $\mathcal{V}_y$ is the target vocabulary. As the parameters of the teacher model are fixed, the training objective can be equivalently written as:

$$\mathcal{J}_{\mathrm{WORD}}(\boldsymbol{\theta}_{x\rightarrow y})$$
$$= -\sum_{\langle \mathbf{x},\mathbf{z}\rangle \in D_{x,z}} \mathbb{E}_{\mathbf{y}|\mathbf{z};\hat{\boldsymbol{\theta}}_{z\rightarrow y}} \Big[S(\mathbf{x}, \mathbf{y}, \mathbf{z}, \hat{\boldsymbol{\theta}}_{z\rightarrow y}, \boldsymbol{\theta}_{x\rightarrow y}) \Big], (11)$$

where

$$S(\mathbf{x}, \mathbf{y}, \mathbf{z}, \hat{\boldsymbol{\theta}}_{z\rightarrow y}, \boldsymbol{\theta}_{x\rightarrow y})$$
$$= \sum_{j=1}^{|\mathbf{y}|} \sum_{y \in \mathcal{V}_y} P(y|\mathbf{z}, \mathbf{y}_{<j}; \hat{\boldsymbol{\theta}}_{z\rightarrow y}) \times$$
$$\log P(y|\mathbf{x}, \mathbf{y}_{<j}; \boldsymbol{\theta}_{x\rightarrow y}). \quad (12)$$

Therefore, our goal is to find a set of source-to-target model parameters that minimizes the training objective:

$$\hat{\boldsymbol{\theta}}_{x\rightarrow y} = \underset{\boldsymbol{\theta}_{x\rightarrow y}}{\operatorname{argmin}} \Big\{ \mathcal{J}_{\mathrm{WORD}}(\boldsymbol{\theta}_{x\rightarrow y}) \Big\}. \quad (13)$$

We use similar approaches as described in Section 3.2 for approximating the full search space with sentence-level teaching. After obtaining $\hat{\boldsymbol{\theta}}_{x\rightarrow y}$, the same decision rule as shown in Equation (1) can be utilized to find the most probable target sentence $\hat{\mathbf{y}}$ for a source sentence $\mathbf{x}$.

4 Experiments

4.1 Setup

We evaluate our approach on the Europarl (Koehn, 2005) and WMT corpora. To compare with pivot-based methods, we use the same dataset as (Cheng et al., 2016a). All the sentences are tokenized by the `tokenize.perl` script. All the experiments treat English as the pivot language and French as the target language.

For the Europarl corpus, we evaluate our proposed methods on Spanish-French (Es-Fr) and German-French (De-Fr) translation tasks in a

Corpus	Direction	Train	Dev.	Test
Europarl	Es→En	850K	2,000	2,000
	De→En	840K	2,000	2,000
	En→Fr	900K	2,000	2,000
WMT	Es→En	6.78M	3,003	3,003
	En→Fr	9.29M	3,003	3,003

Table 1: Data statistics. For the Europarl corpus, we evaluate our approach on Spanish-French (Es-Fr) and German-French (De-Fr) translation tasks. For the WMT corpus, we evaluate approach on the Spanish-French (Es-Fr) translation task. English is used as a pivot language in all experiments.

zero-resource scenario. To avoid the trilingual corpus constituted by the source-pivot and pivot-target corpora, we split the overlapping pivot sentences of the original source-pivot and pivot-target corpora into two equal parts and merge them separately with the non-overlapping parts for each language pair. The development and test sets are from WMT 2006 shared task.[1] The evaluation metric is case-insensitive BLEU (Papineni et al., 2002) as calculated by the `multi-bleu.perl` script. To deal with out-of-vocabulary words, we adopt byte pair encoding (BPE) (Sennrich et al., 2016) to split words into sub-words. The size of sub-words is set to 30K for each language.

For the WMT corpus, we evaluate our approach on a Spanish-French (Es-Fr) translation task with a zero-resource setting. We combine the following corpora to form the Es-En and En-Fr parallel corpora: Common Crawl, News Commentary, Europarl v7 and UN. All the sentences are tokenized by the `tokenize.perl` script. Newstest2011 serves as the development set and Newstest2012 and Newstest2013 serve as test sets. We use case-sensitive BLEU to evaluate translation results. BPE is also used to reduce the vocabulary size. The size of sub-words is set to 43K, 33K, 43K for Spanish, English and French, respectively. See Table 1 for detailed statistics for the Europarl and WMT corpora.

We leverage an open-source NMT toolkit *dl4mt* implemented by Theano [2] for all the experiments and compare our approach with state-of-the-art multilingual methods (Firat et al., 2016b) and pivot-based methods (Cheng et al., 2016a). Two variations of our framework are used in the exper-

iments:

1. Sentence-Level Teaching: for simplicity, we use the mode as suggested in (Kim and Rush, 2016) to approximate the target sentence space in calculating the expected gradients with respect to the expectation in Equation (7). We run beam search on the pivot sentence with the teacher model and choose the highest-scoring target sentence as the mode. Beam size with $k = 1$ (greedy decoding) and $k = 5$ are investigated in our experiments, denoted as *sent-greedy* and *sent-beam*, respectively.[3]

2. Word-Level Teaching: we use the same mode approximation approach as in sentence-level teaching to approximate the expectation in Equation 12, denoted as *word-greedy* (beam search with $k = 1$) and *word-beam* (beam search with $k = 5$) respectively. Besides, Monte Carlo estimation by sampling from the teacher model is also investigated since it introduces more diverse data, denoted as *word-sampling*.

4.2 Assumptions Verification

To verify the assumptions in Section 3.1, we train a source-to-target translation model $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y})$ and a pivot-to-target translation model $P(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}_{z \to y})$ using the trilingual Europarl corpus. Then, we measure the sentence-level and word-level KL divergence from the source-to-target model $P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{x \to y})$ at different iterations to the trained pivot-to-target model $P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z \to y})$ by caculating $\mathcal{J}_{\text{SENT}}$ (Equation (5)) and $\mathcal{J}_{\text{WORD}}$

---

[3]We can also adopt sampling and k-best list for approximation. Random sampling brings a large variance (Sutskever et al., 2014; Ranzato et al., 2015; He et al., 2016) for sentence-level teaching. For k-best list, we renormalize the probabilities

$$P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z \to y}) \sim \frac{P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z \to y})^{\alpha}}{\sum_{\mathbf{y} \in \mathcal{Y}_k} P(\mathbf{y}|\mathbf{z}; \hat{\boldsymbol{\theta}}_{z \to y})^{\alpha}},$$

where $\mathcal{Y}_k$ is the k-best list from beam search of the teacher model and α is a hyperparameter controlling the sharpness of the distribution (Och, 2003). We set $k = 5$ and $\alpha = 5 \times 10^{-3}$. The results on test set for Eureparl Corpus are 32.24 BLEU over Spanish-French translation and 24.91 BLEU over German-French translation, which are slightly better than the sent-beam method. However, considering the traing time and the memory consumption, we think mode approximation is already a good way to approximate the target sentence space for sentence-level teaching.

	Approx.	Iterations				
		0	2w	4w	6w	8w
$\mathcal{J}_{\text{SENT}}$	greedy	313.0	73.1	61.5	56.8	55.1
	beam	323.5	73.1	60.7	55.4	54.0
$\mathcal{J}_{\text{WORD}}$	greedy	274.0	51.5	43.1	39.4	38.8
	beam	288.7	52.7	43.3	39.2	38.4
	sampling	268.6	53.8	46.6	42.8	42.4

Table 2: Verification of sentence-level and word-level assumptions by evaluating approximated KL divergence from the source-to-target model to the pivot-to-target model over training iterations of the source-to-target model. The pivot-to-target model is trained and kept fixed.

	Method	Es→ Fr	De→ Fr
Cheng et al. (2016a)	pivot	29.79	23.70
	hard	29.93	23.88
	soft	30.57	23.79
	likelihood	32.59	25.93
Ours	sent-beam	31.64	24.39
	word-sampling	33.86	27.03

Table 3: Comparison with previous work on Spanish-French and German-French translation tasks from the Europarl corpus. English is treated as the pivot language. The likelihood method uses 100K parallel source-target sentences, which are not available for other methods.

(Equation (9)) on 2,000 parallel source-pivot sentences from the development set of WMT 2006 shared task.

Table 2 shows the results. The source-to-target model is randomly initialized at iteration 0. We find that $\mathcal{J}_{\text{SENT}}$ and $\mathcal{J}_{\text{WORD}}$ decrease over time, suggesting that the source-to-target and pivot-to-target models do have small KL divergence at both sentence and word levels.

4.3 Results on the Europarl Corpus

Table 3 gives BLEU scores on the Europarl corpus of our best performing sentence-level method (sent-beam) and word-level method (word-sampling) compared with pivot-based methods (Cheng et al., 2016a). We use the same data preprocessing as (Cheng et al., 2016a). We find that both the sent-beam and word-sampling methods outperform the pivot-based approaches in a zero-resource scenario across language pairs. Our word-sampling method improves over the best performing zero-resource pivot-based method (soft) on Spanish-French translation by +3.29 BLEU points and German-French translation by +3.24 BLEU points. In addition, the word-sampling mothod surprisingly obtains improvement over the likelihood method, which leverages a source-target parallel corpus. The

Method	Es→ Fr		De→ Fr	
	dev	test	dev	test
sent-greedy	31.00	31.05	22.34	21.88
sent-beam	31.57	31.64	24.95	24.39
word-greedy	31.37	31.92	24.72	25.15
word-beam	30.81	31.21	24.64	24.19
word-sampling	33.65	33.86	26.99	27.03

Table 4: Comparison of our proposed methods on Spanish-French and German-French translation tasks from the Europarl corpus. English is treated as the pivot language.

significant improvements can be explained by the error propagation problem of pivot-based methods that translation error of the source-to-pivot translation process is propagated to the pivot-to-target translation process.

Table 4 shows BLEU scores on the Europarl corpus of our proposed methods. For sentence-level approaches, the sent-beam method outperforms the sent-greedy method by +0.59 BLEU points over Spanish-French translation and +2.51 BLEU points over German-French translation on the test set. The results are in line with our observation in Table 2 that sentence-level KL divergence by beam approximation is smaller than that by greedy approximation. However, as the
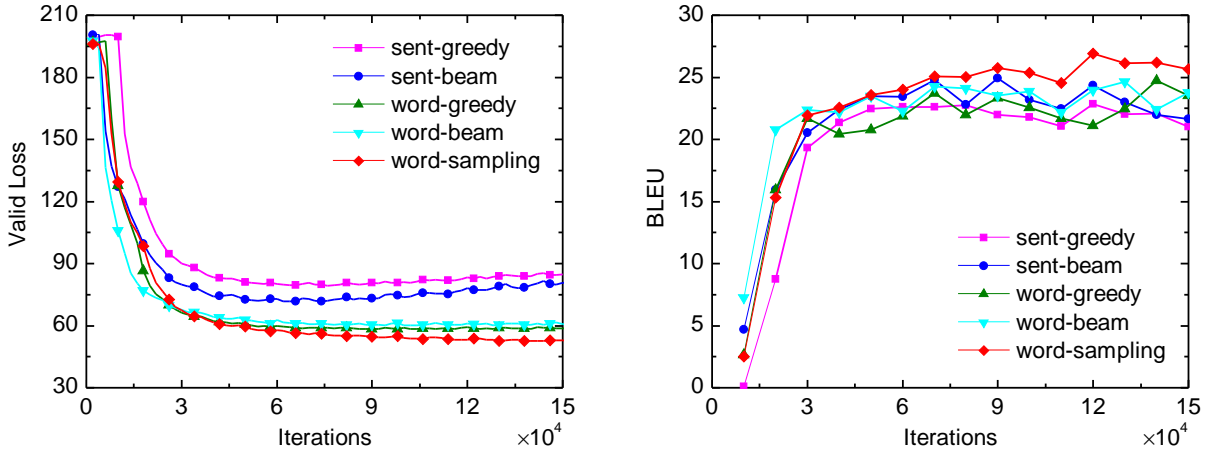
Figure 2: Validation loss and BLEU across iterations of our proposed methods.

	Method	Training			BLEU	
		Es→ En	En→ Fr	Es→ Fr	Newstest2012	Newstest2013
	Existing zero-resource NMT systems					
Cheng et al. (2016a)[†]	pivot	6.78M	9.29M	-	24.60	-
Cheng et al. (2016a)[†]	likelihood	6.78M	9.29M	100K	25.78	-
Firat et al. (2016b)	one-to-one	34.71M	65.77M	-	17.59	17.61
Firat et al. (2016b)[†]	many-to-one	34.71M	65.77M	-	21.33	21.19
	Our zero-resource NMT system					
	word-sampling	6.78M	9.29M	-	28.06	27.03

Table 5: Comparison with previous work on Spanish-French translation in a zero-resource scenario over the WMT corpus. The BLEU scores are case sensitive. †: the method depends on two-step decoding.

time complexity grows linearly with the number of beams k, the better performance is achieved at the expense of search time.

For word-level experiments, we observe that the word-sampling method performs much better than the other two methods: +1.94 BLEU points on Spanish-French translation and +1.88 BLEU points on German-French translation over the word-greedy method; +2.65 BLEU points on Spanish-French translation and +2.84 BLEU points on German-French translation over the word-beam method. Although Table 2 shows that word-level KL divergence approximated by sampling is larger than that by greedy or beam, sampling approximation introduces more data diversity for training, which dominates the effect of KL divergence difference.

We plot validation loss[4] and BLEU scores over iterations on the German-French translation task in Figure 2. We observe that word-level models

tend to have lower validation loss compared with sentence-level methods. Generally, models with lower validation loss tend to have higher BLEU. Our results indicate that this is not necessarily the case: the sent-beam method converges to +0.31 BLEU points on the validation set with +13 validation loss compared with the word-beam method. Kim and Rush (2016) claim a similar observation in data distillation for NMT and provide an explanation that student distributions are more peaked for sentence-level methods. This is indeed the case in our result: on German-French translation task the argmax for the sent-beam student model (on average) approximately accounts for 3.49% of the total probability mass, while the corresponding number is 1.25% for the word-beam student model and 2.60% for the teacher model.

4.4 Results on the WMT Corpus

The word-sampling method obtains the best performance in our five proposed approaches according to experiments on the Europarl corpus. To further verify this approach, we conduct ex-

---

[4]*Validation loss*: the average negative log-likelihood of sentence pairs on the validation set.

groundtruth	source	Os sentáis al volante en la costa oeste , en San Francisco , y vuestra misión es llegar los primeros a Nueva York .
	pivot	You get in the car on the west coast , in San Francisco , and your task is to be the first one to reach New York .
	target	Vous vous asseyez derrière le volant sur la côte ouest à San Francisco et votre mission est d' arriver le premier à New York .
pivot	pivot	You 'll feel at *the west coast in San Francisco , and your* mission *is to* get *the first* to *New York .* [BLEU: 33.93]
	target	*Vous vous* sentirez comme chez vous *à San Francisco* , *et votre mission est d'* obtenir *le premier à New York .* [BLEU: 44.52]
likelihood	pivot	You feel at *the west coast , in San Francisco , and your* mission *is to* reach *the first* to *New York .* [BLEU: 47.22]
	target	*Vous vous* sentez à *la côte ouest* , *à San Francisco* , *et votre mission est d'* atteindre *le premier à New York .* [BLEU: 49.44]
word-sampling	target	*Vous vous* sentez au *volant sur la côte ouest* , *à San Francisco et votre mission est d' arriver le premier à New York .* [BLEU: 78.78]

Table 6: Examples and corresponding sentence BLEU scores of translations using the pivot and likelihood methods in (Cheng et al., 2016a) and the proposed word-sampling method. We observe that our approach generates better translations than the methods in (Cheng et al., 2016a). We italicize *correct translation segments* which are no short than 2-grams.

periments on the large scale WMT corpus for Spanish-French translation. Table 5 shows the results of our word-sampling method in comparison with other state-of-the-art baselines. Cheng et al. (2016a) use the same datasets and the same preprocessing as ours. Firat et al. (2016b) utilize a much larger training set.[5] Our method obtains significant improvement over the pivot baseline by +3.46 BLEU points on Newstest2012 and over many-to-one by +5.84 BLEU points on Newstest2013. Note that both methods depend on a source-pivot-target decoding path. Table 6 shows translation examples of the pivot and likelihood methods proposed in (Cheng et al., 2016a) and our proposed word-sampling method. For the pivot and likelihood methods, the Spainish sentence segment 'sentáis al volante' is lost when translated to English. Therefore, both methods miss this information in the translated French sentence. However, the word-sampling method generates 'volant sur', which partially translates 'sentáis al volante', resulting in improved translation quality of target-language sentence.

4.5 Results with Small Source-Pivot Data

The word-sampling method can also be applied to zero-resource NMT with a small source-pivot corpus. Specifically, the size of the source-pivot corpus is orders of magnitude smaller than that of the pivot-target corpus. This setting makes sense in applications. For example, there are significantly fewer Urdu-English corpora available than

[5]Their training set does not include the Common Crawl corpus.

Method	Corpus			BLEU
	De-En	De-Fr	En-Fr	
MLE	×	√	×	19.30
transfer	×	√	√	22.39
pivot	√	×	√	17.32
Ours	√	×	√	22.95

Table 7: Comparison on German-French translation task from the Europarl corpus with 100K German-English sentences. English is regarded as the pivot language. Transfer represents the transfer learning method in (Zoph et al., 2016). 100K parallel German-French sentences are used for the MLE and transfer methods.

English-French corpora.

To fulfill this task, we combine our best performing word-sampling method with the initialization and parameter freezing strategy proposed in (Zoph et al., 2016). The Europarl corpus is used in the experiments. We set the size of German-English training data to 100K and use the same teacher model trained with 900K English-French sentences.

Table 7 gives the BLEU score of our method on German-French translation compared with three other methods. Note that our task is much harder than transfer learning (Zoph et al., 2016) since it depends on a parallel German-French corpus. Surprisingly, our method outperforms all other methods. We significantly improve the baseline pivot method by +5.63 BLEU points and the state-of-the-art transfer learning method by +0.56 BLEU points.

5 Related Work

Training NMT models in a zero-resource scenario by leveraging other languages has attracted intensive attention in recent years. Firat et al. (2016b) propose an approach which delivers the multi-way, multilingual NMT model proposed by (Firat et al., 2016a) to zero-resource translation. They use the multi-way NMT model trained by other language pairs to generate a pseudo parallel corpus and fine-tune the attention mechanism of the multi-way NMT model to enable zero-resource translation. Several authors propose a universal encoder-decoder network in multilingual scenarios to perform zero-shot learning (Johnson et al., 2016; Ha et al., 2016). This universal model extracts translation knowledge from multiple different languages, making zero-resource translation feasible without direct training.

Besides multilingual NMT, another important line of research is bridging source and target languages via a pivot language. This idea is widely used in SMT (de Gispert and Mariño, 2006; Cohn and Lapata, 2007; Utiyama and Isahara, 2007; Wu and Wang, 2007; Bertoldi et al., 2008; Wu and Wang, 2009; Zahabi et al., 2013; Kholy et al., 2013). Cheng et al. (2016a) propose pivot-based NMT by simultaneously improving source-to-pivot and pivot-to-target translation quality in order to improve source-to-target translation quality. Nakayama and Nishida (2016) achieve zero-resource machine translation by utilizing image as a pivot and training multimodal encoders to share common semantic representation.

Our work is also related to knowledge distillation, which trains a compact model to approximate the function learned by a larger, more complex model or an ensemble of models (Bucila et al., 2006; Ba and Caurana, 2014; Li et al., 2014; Hinton et al., 2015). Kim and Rush (2016) first introduce knowledge distillation in neural machine translation. They suggest to generate a pseudo corpus to train the student network. Compared with their work, we focus on zero-resource learning instead of model compression.

6 Conclusion

In this paper, we propose a novel framework to train the student model without parallel corpora under the guidance of the pre-trained teacher model on a source-pivot parallel corpus. We introduce sentence-level and word-level teaching to guide the learning process of the student model. Experiments on the Europarl and WMT corpora across languages show that our proposed word-level sampling method can significantly outperforms the state-of-the-art pivot-based methods and multilingual methods in terms of translation quality and decoding efficiency.

We also analyze zero-resource translation with small source-pivot data, and combine our word-level sampling method with initialization and parameter freezing suggested by (Zoph et al., 2016). The experiments on the Europarl corpus show that our approach obtains an significant improvement over the pivot-based baseline.

In the future, we plan to test our approach on more diverse language pairs, e.g., zero-resource Uyghur-English translation using Chinese as a pivot. It is also interesting to extend the teacher-student framework to other cross-lingual NLP applications as our method is transparent to architectures.

References

Jimmy Ba and Rich Caurana. 2014. Do deep nets really need to be deep? In *NIPS*.

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

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

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *KDD*.

Yong Cheng, Yang Liu, Qian Yang, Maosong Sun, and Wei Xu. 2016a. Neural machine translation with pivot languages. *CoRR* abs/1611.04928.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016b. Semi-supervised learning for neural machine translation .

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

Adrià de Gispert and José B. Mariño. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*. Citeseer, pages 65–68.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *HLT-NAACL*.

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

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR* abs/1611.04798.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *NIPS*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR* abs/1503.02531.

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

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR* abs/1611.04558.

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

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language independent connectivity strength features for phrase pivot statistical machine translation.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*.

Philipp Koehn. 2005. Europarl: a parallel corpus for statistical machine translation.

Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. 2014. Learning small-size dnn with output-distribution-based criteria. In *INTERSPEECH*.

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

Hideki Nakayama and Noriki Nishida. 2016. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *CoRR* abs/1611.04503.

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

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR* abs/1511.06732.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units .

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks .

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

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation* 21:165–181.

Hua Wu and Haifeng Wang. 2009. Revisiting pivot language approach for machine translation. In *ACL/IJCNLP*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR* abs/1609.08144.

Samira Tofighi Zahabi, Somayeh Bakhshaei, and Shahram Khadivi. 2013. Using context vectors in improving a machine translation system with bridge language. In *ACL*.

Xiaoning Zhu, Zhongjun He, Hua Wu, Haifeng Wang, Conghui Zhu, and Tiejun Zhao. 2013. Improving pivot-based statistical machine translation using random walk. In *EMNLP*.

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