Triangular Architecture for Rare Language Translation

Shuo Ren[1,2]*, Wenhu Chen[3], Shujie Liu[4], Mu Li[4], Ming Zhou[4] and Shuai Ma[1,2]

[1]SKLSDE Lab, Beihang University, Beijing, China
[2]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China
[3]University of California, Santa Barbara, CA, USA
[4]Microsoft Research in Asia, Beijing, China

Abstract

Neural Machine Translation (NMT) performs poor on the low-resource language pair (X, Z), especially when Z is a rare language. By introducing another rich language Y, we propose a novel triangular training architecture (TA-NMT) to leverage bilingual data (Y, Z) (may be small) and (X, Y) (can be rich) to improve the translation performance of low-resource pairs. In this triangular architecture, Z is taken as the intermediate latent variable, and translation models of Z are jointly optimized with a unified bidirectional EM algorithm under the goal of maximizing the translation likelihood of (X, Y). Empirical results demonstrate that our method significantly improves the translation quality of rare languages on MultiUN and IWSLT2012 datasets, and achieves even better performance combining back-translation methods.

1 Introduction

In recent years, Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2014) has achieved remarkable performance on many translation tasks (Jean et al., 2015; Sennrich et al., 2016; Wu et al., 2016; Sennrich et al., 2017). Being an end-to-end architecture, an NMT system first encodes the input sentence into a sequence of real vectors, based on which the decoder generates the target sequence word by word with the attention mechanism (Bahdanau et al., 2014; Luong et al., 2015). During training, NMT systems are optimized to maximize the translation probability of a given language pair

---
*Contribution during internship at MSRA.

with the Maximum Likelihood Estimation (MLE) method, which requires large bilingual data to fit the large parameter space. Without adequate data, which is common especially when it comes to a rare language, NMT usually falls short on low-resource language pairs (Zoph et al., 2016).

In order to deal with the data sparsity problem for NMT, exploiting monolingual data (Sennrich et al., 2015; Zhang and Zong, 2016; Cheng et al., 2016; Zhang et al., 2018; He et al., 2016) is the most common method. With monolingual data, the back-translation method (Sennrich et al., 2015) generates pseudo bilingual sentences with a target-to-source translation model to train the source-to-target one. By extending back-translation, source-to-target and target-to-source translation models can be jointly trained and boost each other (Cheng et al., 2016; Zhang et al., 2018). Similar to joint training (Cheng et al., 2016; Zhang et al., 2018), dual learning (He et al., 2016) designs a reinforcement learning framework to better capitalize on monolingual data and jointly train two models.

Instead of leveraging monolingual data (X or Z) to enrich the low-resource bilingual pair (X, Z), in this paper, we are motivated to introduce another rich language Y, by which additionally acquired bilingual data (Y, Z) and (X, Y) can be exploited to improve the translation performance of (X, Z). This requirement is easy to satisfy, especially when Z is a rare language but X is not. Under this scenario, (X, Y) can be a rich-resource pair and provide much bilingual data, while (Y, Z) would also be a low-resource pair mostly because Z is rare. For example, in the dataset IWSLT2012, there are only 112.6K bilingual sentence pairs of English-Hebrew, since Hebrew is a rare language. If French is introduced as the third language, we can have another low-resource bilingual data of French-Hebrew (116.3K sentence pairs), and easily-acquired bilingual data
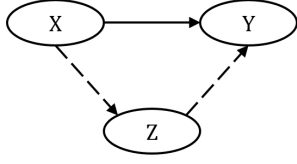
of the rich-resource pair English-French.

Figure 1: Triangular architecture for rare language translation. The solid lines mean rich-resource and the dash lines mean low-resource. X, Y and Z are three different languages.

With the introduced rich language Y, in this paper, we propose a novel triangular architecture (TA-NMT) to exploit the additional bilingual data of (Y, Z) and (X, Y), in order to get better translation performance on the low-resource pair (X, Z), as shown in Figure 1. In this architecture, (Y, Z) is used for training another translation model to score the translation model of (X, Z), while (X, Y) is used to provide large bilingual data with favorable alignment information.

Under the motivation to exploit the rich-resource pair (X, Y), instead of modeling $X \Rightarrow Z$ directly, our method starts from modeling the translation task $X \Rightarrow Y$ while taking Z as a latent variable. Then, we decompose $X \Rightarrow Y$ into two phases for training two translation models of low-resource pairs $((X, Z)$ and $(Y, Z))$ respectively. The first translation model generates a sequence in the hidden space of Z from X, based on which the second one generates the translation in Y. These two models can be optimized jointly with an Expectation Maximization (EM) framework with the goal of maximizing the translation probability $p(y|x)$. In this framework, the two models can boost each other by generating pseudo bilingual data for model training with the weights scored from the other. By reversing the translation direction of $X \Rightarrow Y$, our method can be used to train another two translation models $p(z|y)$ and $p(x|z)$. Therefore, the four translation models $(p(z|x), p(x|z), p(z|y)$ and $p(y|z))$ of the rare language Z can be optimized jointly with our proposed unified bidirectional EM algorithm.

Experimental results on the MultiUN and IWSLT2012 datasets demonstrate that our method can achieve significant improvements for rare languages translation. By incorporating back-translation (a method leveraging more monolingual data) into our method, TA-NMT can achieve even further improvements.

Our contributions are listed as follows:

- We propose a novel triangular training architecture (TA-NMT) to effectively tackle the data sparsity problem for rare languages in NMT with an EM framework.

- Our method can exploit two additional bilingual datasets at both the model and data levels by introducing another rich language.

- Our method is a unified bidirectional EM algorithm, in which four translation models on two low-resource pairs are trained jointly and boost each other.

2 Method

As shown in Figure 1, our method tries to leverage (X, Y) (a rich-resource pair) and (Y, Z) to improve the translation performance of low-resource pair (X, Z), during which translation models of (X, Z) and (Y, Z) can be improved jointly.

Instead of directly modeling the translation probabilities of low-resource pairs, we model the rich-resource pair translation $X \Rightarrow Y$, with the language Z acting as a bridge to connect X and Y. We decompose $X \Rightarrow Y$ into two phases for training two translation models. The first model $p(z|x)$ generates the latent translation in Z from the input sentence in X, based on which the second one $p(y|z)$ generate the final translation in language Y. Following the standard EM procedure (Borman, 2004) and Jensen's inequality, we derive the lower bound of $p(y|x)$ over the whole training data D as follows:

$$
\begin{aligned}
L&(\Theta; D) \\
&= \sum_{(x,y) \in D} \log p(y|x) \\
&= \sum_{(x,y) \in D} \log \sum_z p(z|x) p(y|z) \\
&= \sum_{(x,y) \in D} \log \sum_z Q(z) \frac{p(z|x)p(y|z)}{Q(z)} \qquad (1) \\
&\geq \sum_{(x,y) \in D} \sum_z Q(z) \log \frac{p(z|x)p(y|z)}{Q(z)} \\
&\doteq \mathcal{L}(Q)
\end{aligned}
$$

where Θ is the model parameters set of $p(z|x)$ and $p(y|z)$, and $Q(z)$ is an arbitrary posterior distribution of z. We denote the lower-bound in the last

but one line as $\mathcal{L}(Q)$. Note that we use an approximation that $p(y|x, z) \approx p(y|z)$ due to the semantic equivalence of parallel sentences x and y.

In the following subsections, we will first propose our EM method in subsection 2.1 based on the lower-bound derived above. Next, we will extend our method to two directions and give our unified bidirectional EM training in subsection 2.2. Then, in subsection 2.3, we will discuss more training details of our method and present our algorithm in the form of pseudo codes.

2.1 EM Training

To maximize $L(\Theta; D)$, the EM algorithm can be leveraged to maximize its lower bound $\mathcal{L}(Q)$. In the E-step, we calculate the expectation of the variable z using current estimate for the model, namely find the posterior distribution $Q(z)$. In the M-step, with the expectation $Q(z)$, we maximize the lower bound $\mathcal{L}(Q)$. Note that conditioned on the observed data and current model, the calculation of $Q(z)$ is intractable, so we choose $Q(z) = p(z|x)$ approximately.

M-step: In the M-step, we maximize the lower bound $\mathcal{L}(Q)$ w.r.t model parameters given $Q(z)$. By substituting $Q(z) = p(z|x)$ into $\mathcal{L}(Q)$, we can get the M-step as follows:

$$
\begin{aligned}
\Theta_{y|z} &= \underset{\Theta_{y|z}}{\arg\max} \, \mathcal{L}(Q) \\
&= \underset{\Theta_{y|z}}{\arg\max} \sum_{(x,y)\in D} \sum_z p(z|x) \log p(y|z) \\
&= \underset{\Theta_{y|z}}{\arg\max} \sum_{(x,y)\in D} E_{z\sim p(z|x)} \log p(y|z)
\end{aligned}
$$

$$(2)$$

E-step: The approximate choice of $Q(z)$ brings in a gap between $\mathcal{L}(Q)$ and $L(\Theta; D)$, which can be minimized in the E-step with Generalized EM method (McLachlan and Krishnan, 2007). According to Bishop (2006), we can write this gap explicitly as follows:

$$
\begin{aligned}
L(\Theta; D) - \mathcal{L}(Q) &= \sum_z Q(z) \log \frac{Q(z)}{p(z|y)} \\
&= KL(Q(z)||p(z|y)) \\
&= KL(p(z|x)||p(z|y))
\end{aligned}
$$

$$(3)$$

where $KL(\cdot)$ is the KullbackLeibler divergence, and the approximation that $p(z|x, y) \approx p(z|y)$ is also used above.

In the E-step, we minimize the gap between $\mathcal{L}(Q)$ and $L(\Theta; D)$ as follows:

$$
\Theta_{z|x} = \underset{\Theta_{z|x}}{\arg\min} \, KL(p(z|x)||p(z|y)) \quad (4)
$$

To sum it up, the E-step optimizes the model $p(z|x)$ by minimizing the gap between $\mathcal{L}(Q)$ and $L(\Theta; D)$ to get a better lower bound $\mathcal{L}(Q)$. This lower bound is then maximized in the M-step to optimize the model $p(y|z)$. Given the new model $p(y|z)$, the E-step tries to optimize $p(z|x)$ again to find a new lower bound, with which the M-step is re-performed. This iteration process continues until the models converge, which is guaranteed by the convergence of the EM algorithm.

2.2 Unified Bidirectional Training

The model $p(z|y)$ is used as an approximation of $p(z|x, y)$ in the E-step optimization (Equation 3). Due to the low resource property of the language pair (Y, Z), $p(z|y)$ cannot be well trained. To solve this problem, we can jointly optimize $p(x|z)$ and $p(z|y)$ similarly by maximizing the reverse translation probability $p(x|y)$.

We now give our unified bidirectional generalized EM procedures as follows:

- Direction of $X \Rightarrow Y$

  E: Optimize $\Theta_{z|x}$.

  $$
  \underset{\Theta_{z|x}}{\arg\min} \, KL(p(z|x)||p(z|y)) \quad (5)
  $$

  M: Optimize $\Theta_{y|z}$.

  $$
  \underset{\Theta_{y|z}}{\arg\max} \sum_{(x,y)\in D} E_{z\sim p(z|x)} \log p(y|z) \quad (6)
  $$

- Direction of $Y \Rightarrow X$

  E: Optimize $\Theta_{z|y}$.

  $$
  \underset{\Theta_{z|y}}{\arg\min} \, KL(p(z|y)||p(z|x)) \quad (7)
  $$

  M: Optimize $\Theta_{x|z}$.

  $$
  \underset{\Theta_{x|z}}{\arg\max} \sum_{(x,y)\in D} E_{z\sim p(z|y)} \log p(x|z) \quad (8)
  $$

Based on the above derivation, the whole architecture of our method can be illustrated in Figure 2, where the dash arrows denote the direction of $p(y|x)$, in which $p(z|x)$ and $p(y|z)$ are trained jointly with the help of $p(z|y)$, while the solid ones denote the direction of $p(x|y)$, in which $p(z|y)$ and $p(x|z)$ are trained jointly with the help of $p(z|x)$.
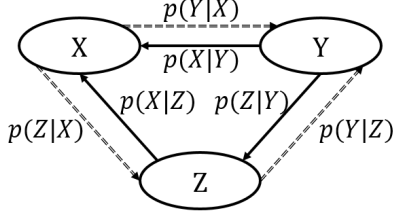
58

Figure 2: Triangular Learning Architecture for Low-Resource NMT

2.3 Training Details

A major difficulty in our unified bidirectional training is the exponential search space of the translation candidates, which could be addressed by either sampling (Shen et al., 2015; Cheng et al., 2016) or mode approximation (Kim and Rush, 2016). In our experiments, we leverage the sampling method and simply generate the top target sentence for approximation.

In order to perform gradient descend training, the parameter gradients for Equations 5 and 7 are formulated as follows:

$$
\begin{aligned}
&\nabla_{\Theta_{z|x}} KL(p(z|x)||p(z|y)) \\
&= E_{z \sim p(z|x)} \log \frac{p(z|x)}{p(z|y)} \nabla_{\Theta_{z|x}} \log p(z|x) \\
&\nabla_{\Theta_{z|y}} KL(p(z|y)||p(z|x)) \\
&= E_{z \sim p(z|y)} \log \frac{p(z|y)}{p(z|x)} \nabla_{\Theta_{z|y}} \log p(z|y)
\end{aligned}
\tag{9}
$$

Similar to reinforcement learning, models $p(z|x)$ and $p(z|y)$ are trained using samples generated by the models themselves. According to our observation, some samples are noisy and detrimental to the training process. One way to tackle this is to filter out the bad ones using some additional metrics (BLEU, etc.). Nevertheless, in our settings, BLEU scores cannot be calculated during training due to the absence of the golden targets (z is generated based on x or y from the rich-resource pair (x, y)). Therefore we choose IBM model1 scores to weight the generated translation candidates, with the word translation probabilities calculated based on the given bilingual data (the low-resource pair (x, z) or (y, z)). Additionally, to stabilize the training process, the pseudo samples generated by model $p(z|x)$ or $p(z|y)$ are mixed with true bilingual samples in the same mini-batch with the ratio of 1-1. The whole training procedure is described in the following Algorithm 1, where the 5th and 9th steps are generating pseudo data.

Algorithm 1 Training low-resource translation models with the triangular architecture

Input: Rich-resource bilingual data (x, y); low-resource bilingual data (x, z) and (y, z)

Output: Parameters $\Theta_{z|x}$, $\Theta_{y|z}$, $\Theta_{z|y}$ and $\Theta_{x|z}$

1: Pre-train $p(z|x)$, $p(z|y)$, $p(x|z)$, $p(y|z)$
2: **while** not convergence **do**
3:      Sample $(x, y), (x^*, z^*), (y^*, z^*) \in D$
4:      $\triangleright X \Rightarrow Y$: Optimize $\Theta_{z|x}$ and $\Theta_{y|z}$
5:      Generate z' from $p(z'|x)$ and build the training batches $B_1 = (x, z') \cup (x^*, z^*)$, $B_2 = (y, z') \cup (y^*, z^*)$
6:      E-step: update $\Theta_{z|x}$ with B_1 (Equation 5)
7:      M-step: update $\Theta_{y|z}$ with B_2 (Equation 6)
8:      $\triangleright Y \Rightarrow X$: Optimize $\Theta_{z|y}$ and $\Theta_{x|z}$
9:      Generate z' from $p(z'|y)$ and build the training batches $B_3 = (y, z') \cup (y^*, z^*)$, $B_4 = (x, z') \cup (x^*, z^*)$
10:      E-step: update $\Theta_{z|y}$ with B_3 (Equation 7)
11:      M-step: update $\Theta_{x|z}$ with B_4 (Equation 8)
12: **end while**
13: **return** $\Theta_{z|x}$, $\Theta_{y|z}$, $\Theta_{z|y}$ and $\Theta_{x|z}$

---

3 Experiments

3.1 Datasets

In order to verify our method, we conduct experiments on two multilingual datasets. The one is MultiUN (Eisele and Chen, 2010), which is a collection of translated documents from the United Nations, and the other is IWSLT2012 (Cettolo et al., 2012), which is a set of multilingual transcriptions of TED talks. As is mentioned in section 1, our method is compatible with methods exploiting monolingual data. So we also find some extra monolingual data of rare languages in both datasets and conduct experiments incorporating back-translation into our method.

MultiUN: English-French (EN-FR) bilingual data are used as the rich-resource pair (X, Y). Arabic (AR) and Spanish (ES) are used as two simulated rare languages Z. We randomly choose subsets of bilingual data of (X, Z) and (Y, Z) in the original dataset to simulate low-resource situations, and make sure there is no overlap in Z between chosen data of (X, Z) and (Y, Z).

IWSLT2012[1]**:** English-French is used as the rich-resource pair (X, Y), and two rare languages Z are Hebrew (HE) and Romanian (RO) in our

---

[1] https://wit3.fbk.eu/mt.php?release=2012-02-plain

59

Pair	MultiUN		IWSLT2012	
	Lang	Size	Lang	Size
(X,Y)	EN-FR	9.9 M	EN-FR [3]	7.9 M
(X,Z)	EN-AR	116 K	EN-HE	112.6 K
(Y,Z)	FR-AR	116 K	FR-HE	116.3 K
mono Z	AR	3 M	HE	512.5 K
(X,Z)	EN-ES	116 K	EN-RO [4]	467.3 K
(Y,Z)	FR-ES	116 K	FR-RO	111.6 K
mono Z	ES	3 M	RO	885.0 K

Table 1: training data size of each language pair.

Method	Resources
PBSMT	$(X,Z), (Y,Z)$
RNNSearch	$(X,Z), (Y,Z)$
T-S	$(X,Z), (Y,Z), (X,Y)$
BackTrans	$(X,Z), (Y,Z), (X,Y)$, mono Z
TA-NMT	$(X,Z), (Y,Z), (X,Y)$
TA-NMT(GI)	$(X,Z), (Y,Z), (X,Y)$, mono Z

Table 2: Resources that different methods use

choice. Note that in this dataset, low-resource pairs (X,Z) and (Y,Z) are severely overlapped in Z. In addition, English-French bilingual data from WMT2014 dataset are also used to enrich the rich-resource pair. We also use additional English-Romanian bilingual data from Europarlv7 dataset (Koehn, 2005). The monolingual data of Z (HE and RO) are taken from the web[2].

In both datasets, all sentences are filtered within the length of 5 to 50 after tokenization. Both the validation and the test sets are 2,000 parallel sentences sampled from the bilingual data, with the left as training data. The size of training data of all language pairs are shown in Table 1.

3.2 Baselines

We compare our method with four baseline systems. The first baseline is the **RNNSearch** model (Bahdanau et al., 2014), which is a sequence-to-sequence model with attention mechanism trained with given small-scale bilingual data. The trained translation models are also used as pre-trained models for our subsequent training processes.

The second baseline is **PBSMT** (Koehn et al., 2003), which is a phrase-based statistical machine translation system. PBSMT is known to perform well on low-resource language pairs, so we want to compare it with our proposed method. And we use the public available implementation of Moses[5] for training and test in our experiments.

The third baseline is a teacher-student alike method (Chen et al., 2017). For the sake of brevity, we will denote it as **T-S**. The process is illustrated in Figure 3. We treat this method as a second baseline because it can also be regarded as a method exploiting (Y,Z) and (X,Y) to improve

the translation of (X,Z) if we regard (X,Z) as the zero-resource pair and $p(x|y)$ as the teacher model when training $p(z|x)$ and $p(x|z)$.

The fourth baseline is back-translation (Sennrich et al., 2015). We will denote it as **BackTrans**. More concretely, to train the model $p(z|x)$, we use extra monolingual Z described in Table 1 to do back-translation; to train the model $p(x|z)$, we use monolingual X taken from (X,Y). Procedures for training $p(z|y)$ and $p(y|z)$ are similar. This method use extra monolingual data of Z compared with our TA-NMT method. But we can incorporate it into our method.
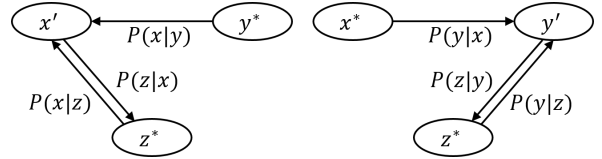

Figure 3: A teacher-student alike method for low-resource translation. For training $p(z|x)$ and $p(x|z)$, we mix the true pair $(y^*, z^*) \in D$ with the pseudo pair (x', z^*) generated by teacher model $p(x'|y^*)$ in the same mini-batch. The training procedure of $p(z|y)$ and $p(y|z)$ is similar.

3.3 Overall Results

Experimental results on both datasets are shown in Table 3 and 4 respectively, in which **RNNSearch**, **PBSMT**, **T-S** and **BackTrans** are four baselines. **TA-NMT** is our proposed method, and **TA-NMT(GI)** is our method incorporating back-translation as good initialization. For the purpose of clarity and a fair comparison, we list the resources that different methods exploit in Table 2.

From Table 3 on MultiUN, the performance of RNNSearch is relatively poor. As is expected, PBSMT performs better than RNNSearch on low-resource pairs by the average of 1.78 BLEU. The T-S method which can doubling the training data

Method	EN2AR (X⇒Z)	AR2EN (Z⇒X)	FR2AR (Y⇒Z)	AR2FR (Z⇒Y)	Ave	EN2ES (X⇒Z)	ES2EN (Z⇒X)	FR2ES (Y⇒Z)	ES2FR (Z⇒Y)	Ave
RNNSearch	18.03	31.40	13.42	22.04	21.22	38.77	36.51	32.92	33.05	35.31
PBSMT	19.44	30.81	15.27	23.65	22.29	38.47	36.64	34.99	33.98	36.02
T-S	19.02	32.47	14.59	23.53	22.40	39.75	38.02	33.67	34.04	36.57
BackTrans	22.19	32.02	15.85	23.57	23.73	42.27	38.42	35.81	34.25	37.76
TA-NMT	20.59	33.22	14.64	24.45	23.23	40.85	39.06	34.52	34.39	37.21
TA-NMT(GI)	**23.16**	**33.64**	**16.50**	**25.07**	**24.59**	**42.63**	**39.53**	**35.87**	**35.21**	**38.31**

Table 3: Test BLEU on MultiUN Dataset.

Method	EN2HE (X⇒Z)	HE2EN (Z⇒X)	FR2HE (Y⇒Z)	HE2FR (Z⇒Y)	Ave	EN2RO (X⇒Z)	RO2EN (Z⇒X)	FR2RO (Y⇒Z)	RO2FR (Z⇒Y)	Ave
RNNSearch	17.94	28.32	11.86	21.67	19.95	31.44	40.63	17.34	25.20	28.65
PBSMT	17.39	28.05	12.77	21.87	20.02	31.51	39.98	18.13	25.47	28.77
T-S	17.97	28.42	12.04	21.99	20.11	31.80	40.86	17.94	25.69	29.07
BackTrans	18.69	28.55	12.31	21.63	20.20	32.18	41.03	18.19	25.30	29.18
TA-NMT	19.19	29.28	12.76	22.62	20.96	33.65	41.93	18.53	26.35	30.12
TA-NMT(GI)	**19.90**	**29.94**	**13.54**	**23.25**	**21.66**	**34.41**	**42.61**	**19.30**	**26.53**	**30.71**

Table 4: Test BLEU on IWSLT Dataset.

for both (X, Z) and (Y, Z) by generating pseudo data from each other, leads up to 1.1 BLEU points improvement on average over RNNSearch. Compared with T-S, our method gains a further improvement of about 0.9 BLEU on average, because our method can better leverage the rich-resource pair (X, Y). With extra large monolingual Z introduced, BackTrans can improve the performance of $p(z|x)$ and $p(z|y)$ significantly compared with all the methods without monolingual Z. However TA-NMT is comparable with or even better than BackTrans for $p(x|z)$ and $p(y|z)$ because both of the methods leverage resources from rich-resource pair (X, Y), but BackTrans does not use the alignment information it provides. Moreover, with back-translation as good initialization, further improvement is achieved by TA-NMT(GI) of about 0.7 BLEU on average over BackTrans.

In Table 4, we can draw the similar conclusion. However, different from MultiUN, in the EN-FR-HE group of IWSLT, (X, Z) and (Y, Z) are severely overlapped in Z. Therefore, T-S cannot improve the performance obviously (only about 0.2 BLEU) on RNNSearch because it fails to essentially double training data via the teacher model. As for EN-FR-RO, with the additionally introduced EN-RO data from Europarlv7, which has no overlap in RO with FR-RO, T-S can improve the average performance more than the EN-FR-HE group. TA-NMT outperforms T-S by 0.93 BLEU on average. Note that even though Back-

Trans uses extra monolingual Z, the improvements are not so obvious as the former dataset, the reason for which we will delve into in the next subsection. Again, with back-translation as good initialization, TA-NMT(GI) can get the best result.

Note that BLEU scores of TA-NMT are lower than BackTrans in the directions of X⇒Z and Y⇒Z. The reason is that the resources used by these two methods are different, as shown in Table 2. To do back translation in two directions (e.g., X⇒Z and Z⇒X), we need monolingual data from both sides (e.g., X and Z), however, in TA-NMT, the monolingual data of Z is not necessary. Therefore, in the translation of X⇒Z or Y⇒Z, Back-Trans uses additional monolingual data of Z while TA-NMT does not, that is why BackTrans outperforms TA-NMT in these directions. Our method can leverage back translation as a good initialization, aka TA-NMT(GI) , and outperforms Back-Trans on all translation directions.

The average test BLEU scores of different methods in each data group (EN-FR-AR, EN-FR-ES, EN-FR-HE, and EN-FR-RO) are listed in the column **Ave** of the tables for clear comparison.

3.4 The Effect of Extra Monolingual Data

Comparing the results of BackTrans and TA-NMT(GI) on both datasets, we notice the improvements of both methods on IWSLT are not as significant as MultiUN. We speculate the reason is the relatively less amount of monolingual Z we use in

the experiments on IWSLT as shown in Table 1. So we conduct the following experiment to verify the conjecture by changing the scale of monolingual Arabic data in the MultiUN dataset, of which the data utilization rates are set to 0%, 10%, 30%, 60% and 100% respectively. Then we compare the performance of BackTrans and TA-NMT(GI) in the EN-FR-AR group. As Figure 4 shows, the amount of monolingual Z actually has a big effect on the results, which can also verify our conjecture above upon the less significant improvement of BackTrans and TA-NMT(GI) on IWSLT. In addition, even with poor "good-initialization", TA-NMT(GI) still get the best results.
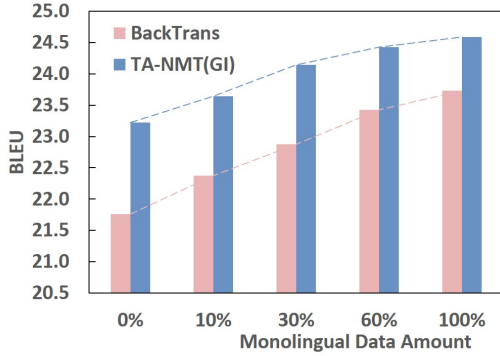

Figure 4: Test BLEU of the EN-FR-AR group performed by BackTrans and TA-NMT(GI) with different amount of monolingual Arabic data.

3.5 EM Training Curves

To better illustrate the behavior of our method, we print the training curves in both the M-steps and E-steps of TA-NMT and TA-NMT(GI) in Figure 5 above. The chosen models printed in this figure are EN2AR and AR2FR on MultiUN, and EN2RO and RO2FR on IWLST.

From Figure 5, we can see that the two low-resource translation models are improved nearly simultaneously along with the training process, which verifies our point that two weak models could boost each other in our EM framework. Notice that at the early stage, the performance of all models stagnates for several iterations, especially of TA-NMT. The reason could be that the pseudo bilingual data and the true training data are heterogeneous, and it may take some time for the models to adapt to a new distribution which both models agree. Compared with TA-NMT, TA-NMT(GI) are more stable, because the models may have
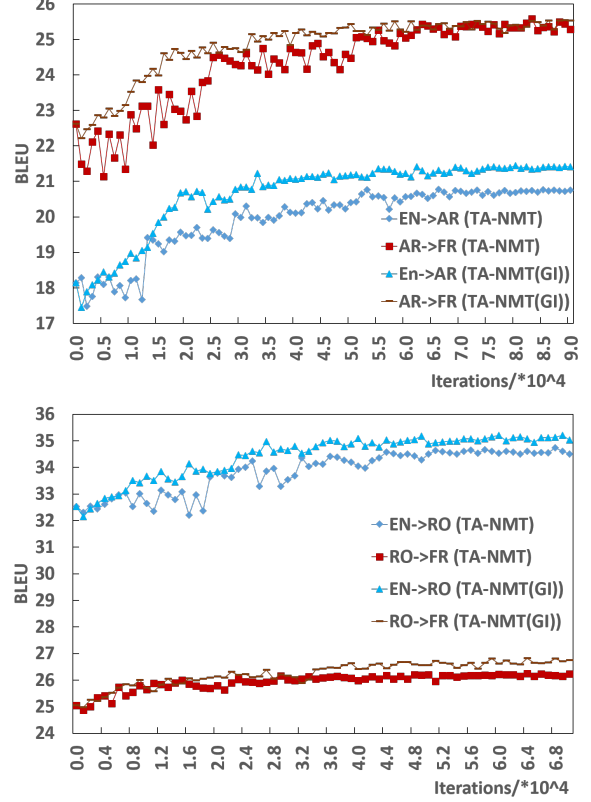

Figure 5: BLEU curves on validation sets during the training processes of TA-NMT and TA-NMT(GI). (Top: EN2AR (the E-step) and AR2FR (the M-step); Bottom: EN2RO (the E-step) and RO2FR (the M-step))

adapted to a mixed distribution of heterogeneous data in the preceding back-translation phase.

3.6 Reinforcement Learning Mechanism in Our Method

As shown in Equation 9, the E-step actually works as a reinforcement learning (RL) mechanism. Models $p(z|x)$ and $p(z|y)$ generate samples by themselves and receive rewards to update their parameters. Note that the reward here is described by the log terms in Equation 9, which is derived from our EM algorithm rather than defined artificially. In Table 5, we do a case study of the EN2ES translation sampled by $p(z|x)$ as well as its time-step rewards during the E-step.

In the first case, the best translation of "political" is "políticos". When the model $p(z|x)$ generates an inaccurate one "políticas", it receives a negative reward (-0.01), with which the model parameters will be updated accordingly. In the sec-

Source	in concluding , poverty eradication requires political will and commitment .
Output	en (0.66) conclusión (0.80) , (0.14) la (0.00) erradicación (1.00) de (0.40) la (0.00) pobreza (0.90) requiere (0.10) voluntad (1.00) y (0.46) compromiso (0.90) políticas (-0.01) . (1.00)
Reference	en conclusión , la erradicación de la pobreza necesita la voluntad y compromiso políticos .
Source	visit us and get to know and love berlin !
Output	visita (0.00) y (0.05) se (0.00) a (0.17) saber (0.00) y (0.04) a (0.01) berlín (0.00) ! (0.00)
Reference	visítanos y llegar a saber y amar a berlín .
Source	legislation also provides an important means of recognizing economic , social and cultural rights at the domestic level .
Output	la (1.00) legislación (0.34) tambin (1.00) constituye (0.60) un (1.00) medio (0.22) importante (0.74) de (0.63) reconocer (0.21) los (0.01) derechos (0.01) econmicos (0.03) , (0.01) sociales (0.02) y (0.01) culturales (1.00) a (0.00) nivel (0.40) nacional (1.00) . (0.03)
Reference	la legislación también constituye un medio importante de reconocer los derechos económicos , iales y culturales a nivel nacional .

Table 5: English to Spanish translation sampled in the E-step as well as its time-step rewards.

ond case, the output misses important words and is not fluent. Rewards received by the model $p(z|x)$ are zero for nearly all tokens in the output, leading to an invalid updating. In the last case, the output sentence is identical to the human reference. The rewards received are nearly all positive and meaningful, thus the RL rule will update the parameters to encourage this translation candidate.

4 Related Work

NMT systems, relying heavily on the availability of large bilingual data, result in poor translation quality for low-resource pairs (Zoph et al., 2016). This low-resource phenomenon has been observed in much preceding work. A very common approach is exploiting monolingual data of both source and target languages (Sennrich et al., 2015; Zhang and Zong, 2016; Cheng et al., 2016; Zhang et al., 2018; He et al., 2016).

As a kind of data augmentation technique, exploiting monolingual data can enrich the training data for low-resource pairs. Sennrich et al. (2015) propose back-translation, exploits the monolingual data of the target side, which is then used to generate pseudo bilingual data via an additional target-to-source translation model. Different from back-translation, Zhang and Zong (2016) propose two approaches to use source-side monolingual data, of which the first is employing a self-learning algorithm to generate pseudo data, while the second is using two NMT models to predict the translation and to reorder the source-side monolingual

sentences. As an extension to these two methods, Cheng et al. (2016) and Zhang et al. (2018) combine two translation directions and propose a training framework to jointly optimize the source-to-target and target-to-source translation models. Similar to joint training, He et al. (2016) propose a dual learning framework with a reinforcement learning mechanism to better leverage monolingual data and make two translation models promote each other. All of these methods are concentrated on exploiting either the monolingual data of the source and target language or both of them.

Our method takes a different angle but is compatible with existing approaches, we propose a novel triangular architecture to leverage two additional language pairs by introducing a third rich language. By combining our method with existing approaches such as back-translation, we can make a further improvement.

Another approach for tackling the low-resource translation problem is multilingual neural machine translation (Firat et al., 2016), where different encoders and decoders for all languages with a shared attention mechanism are trained. This method tends to exploit the network architecture to relate low-resource pairs. Our method is different from it, which is more like a training method rather than network modification.

5 Conclusion

In this paper, we propose a triangular architecture (TA-NMT) to effectively tackle the problem

of low-resource pairs translation with a unified bidirectional EM framework. By introducing another rich language, our method can better exploit the additional language pairs to enrich the original low-resource pair. Compared with the RNNSearch (Bahdanau et al., 2014), a teacher-student alike method (Chen et al., 2017) and the back-translation (Sennrich et al., 2015) on the same data level, our method achieves significant improvement on the MutiUN and IWSLT2012 datasets. Note that our method can be combined with methods exploiting monolingual data for NMT low-resource problem such as back-translation and make further improvements.

In the future, we may extend our architecture to other scenarios, such as totally unsupervised training with no bilingual data for the rare language.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

Sean Borman. 2004. The expectation maximization algorithm-a short tutorial. *Submitted for publication*, pages 1–9.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, volume 261, page 268.

Yun Chen, Yang Liu, Yong Cheng, and Victor OK Li. 2017. A teacher-student framework for zero-resource neural machine translation. *arXiv preprint arXiv:1705.00753*.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. *arXiv preprint arXiv:1606.04596*.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA).

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

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

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for wmt'15. In *WMT@ EMNLP*, pages 134–140.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*, volume 3, page 413.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Geoffrey McLachlan and Thriyambakam Krishnan. 2007. *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural mt systems for wmt17. *arXiv preprint arXiv:1708.00726*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *EMNLP*, pages 1535–1545.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *AAAI*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.