Polygon-Net: A General Framework for Jointly Boosting Multiple Unsupervised Neural Machine Translation Models

Abstract

Neural machine translation (NMT) has achieved great success. However, collecting large-scale parallel data for training is costly and laborious. Recently, unsupervised neural machine translation has attracted more and more attention, due to its demand for monolingual corpus only, which is common and easy to obtain, and its great potentials for the low-resource or even zero-resource machine translation. In this work, we propose a general framework called Polygon-Net, which leverages multi auxiliary languages for jointly boosting unsupervised neural machine translation models. Specifically, we design a novel loss function for multi-language unsupervised neural machine translation. In addition, different from the literature that just updating one or two models individually, Polygon-Net enables multiple unsupervised models in the framework to update in turn and enhance each other for the first time. In this way, multiple unsupervised translation models are associated with each other for training to achieve better performance. Experiments on the benchmark datasets including UN Corpus and WMT show that our approach significantly improves over the two-language based methods, and achieves better performance with more languages introduced to the framework.

1 Introduction

Neural machine translation [Bahdanau et al., 2014; Sutskever et al., 2014] has achieved great success, especially on the majority language pairs [Hassan et al., 2018]. To achieve good performance, large-scale labeled bilingual training corpus are required, since neural translation models usually have large numbers of parameters to be trained. However, these methods would lose their power for the low-resource languages that does not have enough parallel corpus and the zero-resource languages that have no parallel corpus.

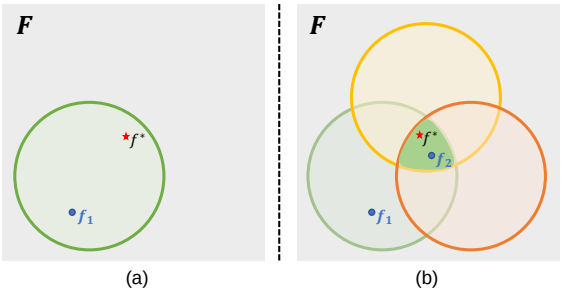

Figure 1: Illustration of the effect of adding more constraints to the training of unsupervised NMT models. F is the full hypothesis space and $f^* \in F$ is the golden translation model. The circles represent different constraints. Left: Constrained only by the duality of Source $\rightarrow$ Target and Target $\rightarrow$ Source objective. some model f_1 which satisfies the constraint may be far from golden translation model f^*. Right: After introducing more constraints, the search space is tightened. As a result, the trained model f_2 in the overlapped space of these constraints could be closer to f^* than f_1.

Notice that monolingual data are very easy to obtain, and even for the low-resource or zero-resource languages, there are rich monolingual corpus on the internet and publications. Therefore, how to leverage monolingual corpus to improve neural machine translation models attracts more and more attention. Recently, the study of unsupervised machine translation [Carbonell et al., 2006; Ravi and Knight, 2011; Firat et al., 2016] provides a new approach for the translation of minority languages.

Core techniques of current top unsupervised NMT systems usually consist of two parts. The first part is to leverage prior knowledge of the two languages to get initial models so as to boost the training process, and the second part is to construct training objectives and principles for the unsupervised learning problem. The challenge is that we cannot explicitly optimize the model like in supervised settings without parallel data. The most popular and effective method is back-translation [Sennrich et al., 2015], which produces pseudo training pairs by mapping sentences in the target language space to the source language space. Then the outputs together with the source sentences forming as pairs are used as parallel data to train the translation model. In short,

back-translation leverages the circuits of Source → Target → Source and Target → Source → Target to design supervisory signals.

The training objectives for unsupervised NMT are usually based on the duality of the Source → Target and Target → Source translation models, while the test is evaluated by the quality of the translations. A potential problem is the gaps between the optimization objectives and the true application scenario. Figure 1(a) illustrate that given the full hypothesis space F, the subspace constrained by the indirect training objectives could be very large. In that case, a fully converged translation model f_1 optimized by the indirect objectives could be far from the golden translation model f^*.[1] Intuitively, if we add more constraints to the search space, it is more easy to search a better f_2 which is closer to f^*.

We try to address this problem from a broader perspective. In addition to the two languages we are focusing on, many other languages also have monolingual corpus. We consider using monolingual corpus from other languages to provide more objectives and improve the performance of the translation models. In particular, We embed languages and translation models into a polygon where each node represents a language and each edge represents a translation model. We introduce new objectives and call the graph Polygon-Net.

Intuitively, a source sentence translated into the same target language along different routes (i.e. different pipelines of translation models) should be (semantically) consistent with each other. Inspired by that, we designed new unsupervised objectives by the probabilistic relations of different path using monolingual corpus other than the original two languages. A technical challenge of computing the objective is that there are infinite probabilities of each routes. We address this challenge by estimating the loss using importance sampling trick. In addition, since the feedback signals among multi-languages are very complicated, we reduce the multi-language situation to some trilingual problems. As a result, the models in the system are updated iteratively, so that multiple models can boost each other, and multiple unsupervised translation systems are combined to train to achieve better performance.

The main contributions of this paper can be summarized as follows: (1) We embed multi languages in to a system called Polygon-Net, and design novel objectives for multi languages using monolingual corpus to jointly boost multiple translation models. (2) To address the challenge of infinite enumeration in computing the objective, we propose to estimate the loss through importance sampling. (3) Experiments on UN Corpus and WMT dataset demonstrates the improvements of our method over baseline models built on two languages.

[1] An extreme case could be the translation mismatch. For example, the word "one" in the source language is translated to "2" in the target language, and the word "2" in the target language is translated to "one" in the source language, which satisfies the training principles but performs bad on test set.

2 Related Work

Our work is related to studies on unsupervised NMT with monolingual data only. Recent top systems includes [Artetxe et al., 2018; Yang et al., 2018; Lample et al., 2018a; Lample et al., 2018b]. They first leverage prior knowledge of the two languages to get initial models which roughly map the source to the target language space. Common techniques includes sharing vocabularies or sub-words of the two languages to get a rough translation model, sharing encoders and decoders, to achieve semantic consistency, and using denoised language model as data-driven prior to the target sentences. After that they construct training objectives and principles for the unsupervised problem using back-translation. Our work is different from them since they all focus on two languages while we propose a method to introduce the monolingual corpus of third-party languages to obtain better models.

There are also some studies on multilingual NMT [Johnson et al., 2016] and design training objectives with auxiliary languages [Ren et al., 2018; Yun et al., 2017]. But they require either parallel data or pre-given translation models, which is different from our setting of using monolingual data only.

3 Model Description

Given a source language space X and a target language space Y, a translation model from language X to Y (denoted by $\mathcal{H}_{XY}$) is usually represented by a conditional distribution $P(y|x; h_{XY})$, where x and y are sentences from language space X and Y respectively, and h_{XY} is the model parameter. Most existing unsupervised neural machine translation methods only consider monolingual corpus of the source and target languages to learn the model. However, we show that the monolingual corpus of third-party languages can be introduced to obtain a better model.

We organize this section as follows. First, we give a brief introduction of the traditional bilingual unsupervised machine translation. Then we show how to introduce only one auxiliary language, i.e. three languages in the system in total, to do the unsupervised machine translation. Specifically, we introduce the objective function induced by multi-path feedback signals, importance sampling estimation for dealing with infinite enumerations, and round training strategy for updating all the models. Finally, we prove that training a system with more than three languages can be simplified to the previously described three-language situation. Based on that, the whole training process of Polygon-Net in a multilingual situation is introduced.

3.1 Traditional Bilingual Unsupervised NMT

The basic idea of traditional bilingual unsupervised NMT is that, if we input a sentence x to the translator from language X to Y along with the translator from language Y to X in the pipeline, the output should be (semantically) consistent with the original sentence x.

Based on the above intuition, [Lample et al., 2018b] design a specific bilingual loss, which consists of two parts.

The first part is back-translation loss, defined in Eqn. (1), where $\mathcal{H}^*_{XY}(x) = \arg\max_y P(y|x; h_{XY})$ and $\mathcal{H}^*_{YX}(y) = \arg\max_x P(x|y; h_{YX})$. Back-translation loss characterizes the disagreement between the original sentence and the back-translation sentence. Note that unlike parameter h_{XY}, the order of subscript in loss does not matter.

$$\begin{aligned}\mathcal{L}^{back}_{XY} &= \mathbb{E}_{y\sim Y}[-\log P(y|\mathcal{H}^*_{YX}(y); h_{XY})] \\ &+ \mathbb{E}_{x\sim X}[-\log P(x|\mathcal{H}^*_{XY}(x); h_{YX})]. \end{aligned} \quad (1)$$

Language model loss is the other part of bilingual loss, which is applied to work as data-driven prior to the target sentences:

$$\begin{aligned}\mathcal{L}^{lm}_{XY} &= \mathbb{E}_{x\sim X}[-\log P(x|M(x); h_{XX})] \\ &+ \mathbb{E}_{y\sim Y}[-\log P(y|M(y); h_{YY})], \end{aligned} \quad (2)$$

where M is a pre-determined noise model with some words dropped and swapped, and h_{XX} (h_{YY}) is the parameter of model which combines h_{XY}'s (h_{YX}'s) encoder and h_{YX}'s (h_{XY}'s) decoder. Then the final bilingual loss is defined as

$$\mathcal{L}^{bi}_{XY} = \mathcal{L}^{back} + \beta\mathcal{L}^{lm}, \quad (3)$$

where β is a hyperparameter controlling the tradeoff between the back-translation loss and the language modeling loss.

3.2 Trilingual Training

We now introduce the auxiliary languages to help doing the unsupervised NMT. We first study the simplest case (only using one auxiliary languages) in this section, and leave the multi-auxiliary-language case in Section 3.3.

Suppose that X and Y are the languages of our interests, and the auxiliary language is denoted by Z. Let us consider two semantic paths: $Z \to Y$ and $Z \to X \to Y$ as shown in Figure 2. Intuitively, a sentence translated into the same final language through different paths should get results (semantically) consistent with each other. In particular, for each source sentence $z \in Z$ and each target sentence $y \in Y$, we have

$$P(y|z; h_{ZY}) = \sum_{x\in X} P(x|z; h_{ZX})P(y|x; h_{XY}). \quad (4)$$

In practice, the above relation should approximately hold for good translation models. Thus we can define the following loss.

$$\begin{aligned}\mathcal{L}^{tri(1)\prime}_{Z} = \sum_{z\in Z} \mathbb{E}_{y\sim P(y|z;h_{ZY})}\Big[&\log P(y|z; h_{ZY}) \\ &- \log \sum_{x\in X} P(x|z; h_{ZX})P(y|x; h_{XY})\Big]^2. \end{aligned} \quad (5)$$

We remark that the expectation of y can be approximately calculated by taking the average of one or more random sampled sentences from distribution $P(y|z; h_{ZY})$.

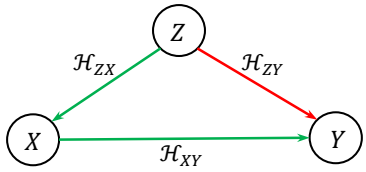

Figure 2: An example of a trilingual translation system. The nodes represent languages and the directed egdes represent translation models. By introducing Z as an auxiliary language, We design a trilingual loss motivated by the probabilistic relationship (see Eqn. (4)) between translation models along the two path marked by the green and red arrows. With that training objective, model $\mathcal{H}_{XY}$, $\mathcal{H}_{ZY}$, and $\mathcal{H}_{ZX}$ are updated and improved jointly.

Importance Sampling

Since language space X is too large, it is impossible to directly compute the sum of $P(x|z; h_{ZX})P(y|x; h_{XY})$ in Eqn. (5). A naive way to address this issue is to build an approximate estimator by sampling enough sentences from the entire language space. However, for most $x \in X$, the value of $P(x|z; h_{ZX})$ and $P(y|x; h_{XY})$ would be almost zero. This is because, only a few sentences from Y or Z are semantically similar to a certain sentence x. To overcome the above issue, we make an identity transform as follows:

$$\begin{aligned}&\sum_x P(x|z; h_{ZX})P(y|x; h_{XY}) \\ &= \sum_x \frac{P(x|z; h_{ZX})P(y|x; h_{XY})P(x|y; h_{YX})}{P(x|y; h_{YX})} \\ &= \mathbb{E}_{x\sim P(x|y;h_{YX})}\frac{P(x|z; h_{ZX})P(y|x; h_{XY})}{P(x|y; h_{YX})} \quad (6) \\ &\approx \frac{1}{K}\sum_{i=1}^{K}\frac{P(x_i|z; h_{ZX})P(y|x_i; h_{XY})}{P(x_i|y; h_{YX})}, \quad (7)\end{aligned}$$

where K is the number of samples and x_i is the sample drawn from distribution $P(x|y; h_{YX})$. In this way, the expectation in Eqn. (6) can be approximately calculated by taking the average of samples's values as Eqn. (7). In other words, we use model $\mathcal{H}_{YX}$ to sample sentences in X which are strongly related to y and z semantically, such that the probability value of $P(x|z; h_{ZX})P(y|x; f)$ in Eqn. (6) is large enough.

This procedure is exactly the technique of importance sampling [Hesterberg, 1988; Hesterberg, 1995]. Therefore, the loss function in Eqn. (5) can be revised as

$$\begin{aligned}\mathcal{L}^{tri(1)\prime}_{Z} &\approx \sum_{z\in Z} \mathbb{E}_{y\sim P(y|z;h_{ZY})}\Big[\log P(y|z; h_{ZY}) \\ &\quad - \log \frac{1}{K}\sum_{i=1}^{K}\frac{P(x_i|z; h_{ZX})P(y|x_i; h_{XY})}{P(x_i|y; h_{YX})}\Big]^2 \\ &=: \mathcal{L}^{tri(1)}_{Z},\end{aligned}$$

where x_i is sampled from distribution $P(x|y; h_{YX})$.

We remark that, in the training process, one can compute $\nabla_{h_{XY}}\mathcal{L}_Z^{tri(1)}, \nabla_{h_{ZX}}\mathcal{L}_Z^{tri(1)}$ and $\nabla_{h_{ZY}}\mathcal{L}_Z^{tri(1)}$ respectively and update models h_{XY}, h_{ZX} and h_{ZY} iteratively to make the training process more stable.

Similarly, we can define the symmetrical loss as

$$\mathcal{L}_Z^{tri(2)} = \sum_{z\in Z} \mathbb{E}_{x\sim P(x|z;h_{ZX})}\Big[\log P(x|z;h_{ZX}) - \log\frac{1}{K}\sum_{i=1}^{L}\frac{P(y_i|z;h_{ZY})P(x|y_i;h_{YX})}{P(y_i|x;h_{XY})}\Big]^2,$$

where y_i is sampled from distribution $P(y|x;h_{XY})$, and K is also the number of samples.

The total loss of using Z as the auxiliary language combines the two parts of loss together:

$$\mathcal{L}_Z^{tri} = \mathcal{L}_Z^{tri(1)} + \mathcal{L}_Z^{tri(2)}. \quad (8)$$

In the same way, we can also define $\mathcal{L}_X^{tri}$ and $\mathcal{L}_Y^{tri}$.

Round Training Strategy

With the help of the auxiliary language, the unsupervised learning feedback signal is added to improve the performance of $\mathcal{H}_{XY}$ and $\mathcal{H}_{YX}$. In the meanwhile, $\mathcal{H}_{ZX}$ and $\mathcal{H}_{ZY}$ are also improved. Here we propose round training strategy, which takes X, Y and Z as the auxiliary language in turns and iteratively updates all the models in the system, so that the models can help each other to improve the performance.

The final objective function for trilingual system is

$$\mathcal{L}_{X,Y,Z} = \mathcal{L}_{XY}^{bi} + \mathcal{L}_{YZ}^{bi} + \mathcal{L}_{XZ}^{bi} + \gamma(\mathcal{L}_X^{tri} + \mathcal{L}_Y^{tri} + \mathcal{L}_Z^{tri}), \quad (9)$$

where γ is used to balance bilingual and trilingual loss.

3.3 Polygon-Net: Extension to More Languages

In this subsection, we extend one auxiliary language to multi auxiliary languages to improve the unsupervised NMT models. We construct a directed graph to illustrate our framework (called Polygon-Net).

As shown in Figure 3 (a), each node represents a language and each directed edge represents a translation model from the starting point to the ending point. By the same idea of Eqn. (4), given a source sentence $z_i \in Z_i$ and a target sentence $z_j \in Z_j$ where Z_i and Z_j are two languages, the translations through two different paths should have the same result as shown in Figure 3 (b). Thus we could define a loss to characterize the difference between every two paths like Eqn. (5). However, the computational complexity would be high. In particular, if we take sentence $z_1 \in Z_1$ as the source sentence and $z_n \in Z_n$ as the target sentence, then the probability of getting z_n from z_1 through any semantic path $Z_1 \to Z_2 \to \cdots \to Z_n$ is calculated by

$$P(z_n|z_1; h_{Z_{n-1}Z_n}, \ldots, h_{Z_1Z_2}) = \sum_{z_2\in Z_2}\cdots\sum_{z_{n-1}\in Z_{n-1}}\prod_{i=1}^{n-1} P(z_{i+1}|z_i; h_{Z_iZ_{i+1}}).$$

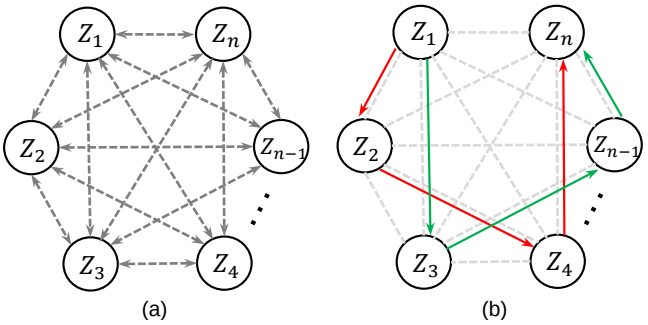

Figure 3: (a) Polygon-Net framework for jointly boosting multiple unsupervised NMT models. Nodes represent languages and directed edges represent directed translation models. Given monolingual corpus in n languages, Polygon-Net enables the models to update according to the objective designed by the probabilistic relations along semantic paths. (b) Two possible semantic paths from Z_1 to Z_n marked by the red and green arrows. Ideally, the translation results along all possible paths from Z_1 to Z_n should be identical in semantics.

Algorithm 1: Multilingual Training Process for Polygon-Net

Require Monolingual corpus of n languages $Z_1, \cdots, Z_n$, sample size K.
repeat
  Sample three languages X, Y, and Z from $\{Z_1, Z_2, \cdots, Z_n\}$, and get three mini-batches of monolingual sentences from each language;
  For each sentence y in the mini-batch of language Y, sample K sentences $\hat{x_1}, \hat{x_2}, \cdots, \hat{x_K}$ according to translation model h_{YX};
  For each sentence x in the mini-batch of language X, sample K sentences $\hat{y_1}, \hat{y_2}, \cdots, \hat{y_K}$ according to translation model h_{XY};
  Calculate the training objective $\mathcal{L}_{\mathcal{XYZ}}$ according to Eqn. (9) and compute the gradient of $\mathcal{L}_{\mathcal{XYZ}}$ with respect to h_{XY},h_{YX},h_{ZX} and h_{ZY} respectively;
  Update the models:
    $h_{XY} \leftarrow h_{XY} - \alpha\nabla_{h_{XY}}\mathcal{L}_{\mathcal{XYZ}}$
    $h_{YX} \leftarrow h_{YX} - \alpha\nabla_{h_{YX}}\mathcal{L}_{\mathcal{XYZ}}$
    $h_{ZX} \leftarrow h_{ZX} - \alpha\nabla_{h_{ZX}}\mathcal{L}_{\mathcal{XYZ}}$
    $h_{ZY} \leftarrow h_{ZY} - \alpha\nabla_{h_{ZY}}\mathcal{L}_{\mathcal{XYZ}}$
until model converges;

The computational complexity is $O(\prod_{i=2}^{n-1}|Z_i|)$ where $|\cdot|$ is the cardinality. Moreover, the computational complexity increases exponentially with the number of nodes.

Instead, we can sum up the loss for each language triplet to be the loss for Polygon-Net as in Figure (4). There are several advantages for the above loss. First, Section 3.2 can be directly applied to calculate the loss.

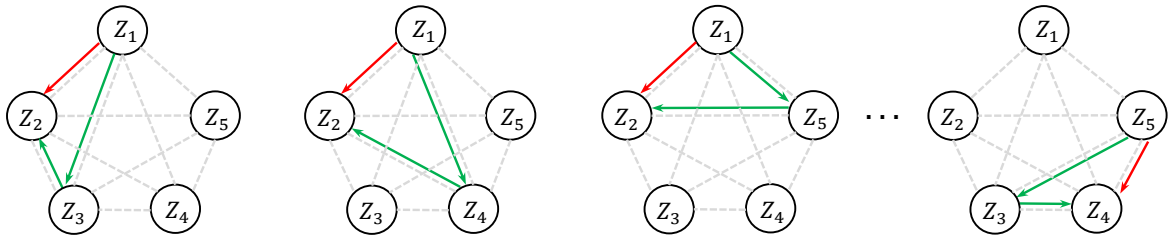

Figure 4: Examples of different trilingual objectives that decomposed from Polygon-Net for five languages.

Model	UN Corpus						WMT					
	en→fr	fr→en	en→es	es→en	fr→es	es→fr	en→fr	fr→en	en→de	de→en	fr→de	de→fr
[Artetxe et al., 2018]	25.32	26.24	35.42	37.82	26.55	25.46	12.95	8.28	7.89	9.94	5.27	3.12
[Lample et al., 2018a]	25.43	26.32	37.56	36.78	26.32	25.86	12.70	7.16	8.32	10.20	4.10	3.03
[Lample et al., 2018b]	35.49	35.86	45.55	45.09	37.12	36.18	18.83	18.68	12.24	15.20	8.86	4.06
Polygon-Net	37.26	37.63	46.44	45.78	38.27	37.32	19.27	19.25	13.05	15.90	10.57	6.05

Table 1: BLEU scores on UN corpus and WMT respectively. Polygon-Net is trained on three languages.

Second, the computational complexity is much lower. Third, if the above loss is zero, then the loss for any two different paths will also be zero. For example, let us consider two paths: $Z_1 \to Z_2 \to \cdots \to Z_n$ and $Z_1 \to Z_n$. We have

$$
\begin{aligned}
&P(z_n|z_1; h_{Z_1 Z_n}) \\
&= \sum_{z_2 \in Z_2} P(z_n|z_2; h_{Z_2 Z_n}) P(z_2|z_1; h_{Z_1 Z_2}) \\
&= \sum_{z_2 \in Z_2} \sum_{z_3 \in Z_3} P(z_n|z_3; h_{Z_3 Z_n}) P(z_3|z_2; h_{Z_2 Z_3}) P(z_2|z_1; h_{Z_1 Z_2}) \\
&= \sum_{z_2 \in Z_2} \cdots \sum_{z_{n-1} \in Z_{n-1}} \prod_{i=1}^{n-1} P(z_{i+1}|z_i; h_{Z_i Z_{i+1}}) \\
&= P(z_n|z_1; h_{Z_{n-1} Z_n}, \ldots, h_{Z_1 Z_2}),
\end{aligned} \tag{10}
$$

where each equality holds because the loss for each triplet is zero. This means that optimizing the loss for triplets could help optimize the loss for every two paths. Therefore, we use the triplet loss for the Polygon-Net. The final objective function for multilingual system is

$$
\mathcal{L}_{Z_1, Z_2, \cdots, Z_n} = \sum_{i=1}^{n} \sum_{j<i} \mathcal{L}^{bi}_{Z_i Z_j} + \gamma \sum_{i \in [n]} \mathcal{L}^{tri}_{Z_i}, \tag{11}
$$

where γ is a coefficient to balance bilingual loss and trilingual loss.

4 Experimental Results

We conduct a serious of experiments to evaluate our Polygon-Net framework. In this section, we first describe the experimental setups and baseline models. And then, to prove the effectiveness of our method, we apply Polygon-Net framework to the three-language situation and compare the performance with another unsupervised baselines on two datasets. After that, we apply our proposed method to four and five languages to evaluate its generalization ability to larger systems, and analyze the impact of introducing more languages. At last, we make some studies and discussions on the selection of hyper-parameters and time cost.

4.1 Settings and Baselines

We compare our proposed Polygon-Net framework with other recent work on unsupervised NMT.

[Artetxe et al., 2018] Unsupervised embedding mappings based encoder-decoder NMT system with denoising and back-translation..

[Lample et al., 2018a] The method takes sentences from monolingual corpora in two different languages and maps them into one common latent space. By learning to reconstruct in both languages from the shared latent feature space, the model effectively learns to translate without using any labeled data.

[Lample et al., 2018b] The model leverages a careful initialization of the parameters, the denoising effect of language models, iterative back-translation and sharing latent representations from two different languages.

Note that the work of [Lample et al., 2018b] is a very strong baseline, which can achieve about 5 to 10 BLUE [Papineni et al., 2002] scores improvement over [Artetxe et al., 2018; Yang et al., 2018; Lample et al., 2018a] on English, French and German language pairs on WMT.

To ensure fair comparison and convincing evaluation, we follow the settings of [Lample et al., 2018b] in our experiments. Specifically, we use Moses scripts [Koehn et al., 2007] for tokenization. We pre-processed the corpus with byte pair encoding (BPE) [Sennrich et al., 2016] since it has been proved to be an effective approach to handle the large vocabulary issue in NMT. We use NMT models of [Lample et al., 2018b] built on transformer [Vaswani et al., 2017] cells. To be fair, the encoders and decoders of our Polygon-Net and baseline pair-wise unsupervised NMT models are all equipped

with 6-layer Transformer with word embedding size 512 and hidden state size 512. The sentences come from different language are distinguished with a language identifier at first token. Before feeding corpus to the NMT models, we first learn BPE embeddings from monolingual corpora in different languages jointly by fastText [Bojanowski et al., 2017], the joint BPE embeddings are used for initializing NMT models. All encoders and decoders are shared across the two languages. Adam optimizer with an initial learning rate 10^{-4}, $\beta_1 = 0.5$, and a batch size of 32 is applied to all experiments in this paper.

4.2 Polygon-Net on Three Languages

In order to verify our proposed multilingual machine translation training framework, we conduct experiments on two different multilingual datasets:

UN Corpus [Ziemski et al., 2016] Parallel corpus composed of United Nations documents. We consider three language pairs: English-French, English-Spanish, French-Spanish. We randomly shuffle the UN corpus and then sample 10 million sentences to construct monolingual corpus of English (en), French (fr), and Spanish (es) respectively. Standard test sets of UN corpus are used for evaluation.

WMT [Bojar et al., 2017] We consider three language pairs: English-French, English-German, French-German. For each language of English (en), French (fr), and German (de), we randomly selected 10 million sentences from WMT monolingual News Crawl datasets from years 2007 through 2010. Since there is no test set for French-German after 2013, we use the test set of WMT'13 for evaluation. Considering [Lample et al., 2018b] reported their results on newstest 2014 for $en-fr$, and newstest 2016 for $en-de$, to ensure the credibility of our reproduced baseline [Lample et al., 2018b], we compared our results with the original reported results on these test sets, which is 17.5 (our implementation) vs. 17.2 (reported) for $en \rightarrow de$, 21.2 vs. 21.0 for $de \rightarrow en$, 24.2 vs. 24.2 for $fr \rightarrow en$ and 24.8 vs. 25.1 for $en \rightarrow fr$.

In Table 1, we report the performance of different unsupervised NMT methods. From the figure we can observe that Polygon-Net outperforms all the baseline models on all language pairs on two datasets. In most of the 12 groups of results, Polygen-Net achieves more than 1 point improvement over the strongest bilingual baseline. The results show that with the help of introducing auxiliary languages, Polygon-Net perform better than methods trained with objectives of two languages, which agrees with our intuition that the constraints from other languages can enhance the performance of target bilingual translation.

4.3 Expand to More Languages

After demonstrating the effectiveness of Polygon-Net on trilingual NMT task, we try to introduce more languages to prove the generality of Polygon-Net further. We consider the other two monolingual corpus in Spanish (es) and Czech (cs) from WMT News Crawl datasets. For Spanish, the News Crawl dataset from 2007 through 2010 only consists of 3.5 million sentences, thus we augment Spanish corpus with the monolingual dataset from the News Crawl dataset from 2007 through 2012, resulting in 10 million sentences for each language. Similarly, the multilingual-aligned test sets from WMT'12 and WMT'13 are used for evaluation.

We report the impact of introducing 3/4/5 languages to the system in Table 2. The results show that Polygon-Net achieves better results with more languages added to the system, which matches our intuition that providing more constraints and objectives could help improve the performance.

The complete results of five-language Polygon-Net is listed in Table 3. Polygon-Net achieves improvement on all language pairs over the strong bilingual baseline. Another interesting observation is that Polygon-Net is helpful to enhance not only the performance of NMT models among non-mainstream languages like $es - cs$, but also can improve models among mainstream languages like $en - fr$ thanks to the constraints from non-mainstream languages.

4.4 Discussions

Impact of Sample Size K We discuss the selection of sample size K. We conduct experiments on UN Corpus and report the results in Table 4. By increasing K, Polygon-Net achieves more accurate estimation of Eqn. (6), resulting in better performance. To balance the performance and computation cost, we set k as 2 for all the experiments in this paper

Time Cost The mainly extra time cost of Polygon-Net compared to bilingual baseline [Lample et al., 2018b] comes from importance sampling to draw sentences from translation models. However, since we introduce more language to tighten the hypothesis space, according to our experience, all the models in the system can converge with fewer iterations. Empirically, the overall time cost of the three-language Polygon-Net is about 1.5 times of bilingual baseline. And the time cost of five-language Polygon-Net is about 2 times of bilingual baseline. Therefore, the time cost is practicable for training multilingual MT systems.

5 Conclusion and Future Work

We proposed Polygon-Net framework, which leverages multi auxiliary languages for jointly boosting unsupervised neural machine translation models. Experimental results verified the effectiveness of our method.

In the future, we will study how to generalize/apply our Polygon-Net to supervised and semi-supervised machine translation.

Model	fr→en	en→fr	de→en	en→de	fr→de	de→fr
[Lample et al., 2018b]	18.83	18.68	12.24	15.20	8.86	4.06
Polygon-Net $[en, fr, de]$	19.27	19.25	13.05	15.90	10.57	6.05
Polygon-Net $[en, fr, de, es]$	20.09	19.61	13.14	16.27	10.82	9.06
Polygon-Net $[en, fr, de, es, cs]$	20.27	19.87	13.25	16.46	11.24	9.45

Table 2: Performance of introducing more languages to Polygon-Net. Tested on WMT'13.

	On WMT'12 Test Set		On WMT'13 Test Set	
	Baseline	Polygon-Net	Baseline	Polygon-Net
fr→en	18.58	19.65	18.68	19.87
en→fr	17.47	19.46	18.83	20.27
de→en	14.61	15.46	15.20	16.46
en→de	11.39	11.97	12.24	13.25
es→en	22.58	23.05	19.67	20.35
en→es	22.86	23.89	19.30	20.43
cs→en	10.88	11.31	10.10	11.56
en→cs	5.99	6.33	6.06	6.79
de→fr	4.00	8.78	4.06	9.45
fr→de	9.35	11.08	8.86	11.24
es→fr	25.95	26.30	23.74	24.13
fr→es	26.12	26.69	23.23	23.88
cs→fr	10.29	11.77	9.34	10.61
fr→cs	7.37	7.93	7.22	8.89
es→de	10.47	11.49	10.07	11.20
de→es	13.63	14.76	12.60	13.68
cs→de	10.72	11.33	10.51	11.16
de→cs	9.37	9.74	10.06	10.52
cs→es	11.43	12.20	10.34	11.29
es→cs	8.10	8.76	6.94	8.32

Table 3: Comparison of BLEU scores on WMT of bilingual baseline [Lample et al., 2018b] and multilingual Polygon-Net.

K	1	2	3	4	5
BLEU	36.32	37.26	37.28	37.29	37.32

Table 4: Impact of K for trilingual Polygon-Net on $en \rightarrow fr$ test set of UN corpus.

References

[Artetxe et al., 2018] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In Proceedings of the Sixth International Conference on Learning Representations, April 2018.

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

[Bojanowski et al., 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[Bojar et al., 2017] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. Findings of the 2017 conference on machine translation (wmt17). In Proceedings of the Second Conference on Machine Translation, pages 169–214, 2017.

[Carbonell et al., 2006] Jaime G Carbonell, Steve Klein, David Miller, Mike Steinbaum, Tomer Grassiany, and Jochen Frey. Context-based machine translation. 2006.

[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. arXiv preprint arXiv:1606.04164, 2016.

[Hassan et al., 2018] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. Achieving human parity on automatic chinese to english news translation. arXiv preprint arXiv:1803.05567, 2018.

[Hesterberg, 1988] Timothy Classen Hesterberg. Advances in importance sampling. PhD thesis, Stanford University, 1988.

[Hesterberg, 1995] Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. Technometrics, 37(2):185–194, 1995.

[Johnson et al., 2016] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. 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, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pages 177–180. Association for Computational Linguistics, 2007.

[Lample et al., 2018a] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In Proceedings of the Sixth International Conference on Learning Representations, April 2018.

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

[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 the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics, 2002.

[Ravi and Knight, 2011] Sujith Ravi and Kevin Knight. Deciphering foreign language. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 12–21, 2011.

[Ren et al., 2018] Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, Ming Zhou, and Shuai Ma. Triangular architecture for rare language translation. 2018.

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

[Sennrich et al., 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725. Association for Computational Linguistics, 2016.

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

[Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems 30, pages 5998–6008. 2017.

[Yang et al., 2018] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing. arXiv preprint arXiv:1804.09057, 2018.

[Yun et al., 2017] Chen Yun, Liu Yang, Cheng Yong, and Victor O. K. Li. A teacher-student framework for zero-resource neural machine translation. 2017.

[Ziemski et al., 2016] Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1. 0. In LREC, 2016.