Semi-Supervised Neural Machine Translation via Marginal Distribution Estimation

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Enhong Chen, and Tie-Yan Liu

***Abstract*—Neural machine translation (NMT) heavily relies on parallel bilingual corpora for training. Since large-scale, high-quality parallel corpora are usually costly to collect, it is appealing to exploit monolingual corpora to improve NMT. Inspired by the law of total probability, which connects the probability of a given target-side monolingual sentence to the conditional probability of translating from a source sentence to the target one, we propose to explicitly exploit this connection and help the training procedure of NMT models using monolingual data. The key technical challenge of this approach is that there are exponentially many source sentences for a target monolingual sentence while computing the sum of the conditional probability given each possible source sentence. We address this challenge by leveraging the reverse translation model (target-to-source translation model) to sample several mostly likely source-side sentences and avoid enumerating all possible candidate source sentences. Then we propose two different methods to leverage the law of total probability, including marginal distribution regularization and likelihood maximization of monolingual corpora. Experiment results on English→French and German→English tasks demonstrate that our methods achieve significant improvement over several strong baselines.**

***Index Terms*—Neural machine translation, semi-supervised learning, natural language processing.**

I. Introduction

MACHINE translation aims at mapping a sentence from the source language space $\mathcal{X}$ into the target language space $\mathcal{Y}$. Recent development of neural networks has witnessed the success of Neural Machine Translation (NMT), which has achieved state-of-the-art performance [1]–[3] through end-to-end learning. In particular, given a parallel sentence pair (x, y), where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the learning objective of most NMT algorithms is to maximize the conditional probability $P_\theta(y|x)$ parameterized by θ.

While neural networks have led to better performance, the huge number, usually tens of millions, of parameters in the NMT model raises a major challenge that it heavily relies on large-scale parallel bilingual corpora for model training. Unfortunately, it is usually quite difficult to collect adequate high-quality parallel corpora. To address this challenge, increasing attention has been paid to leveraging other more easily obtained information, especially huge amount of monolingual corpora on the web, to improve NMT.

Early works [4] proposed to train language models [5], [6] independently with target-side monolingual sentences, and incorporate them into NMT models during decoding by re-scoring the candidate words according to the weighted sum of the scores provided by the translation model and the language model, or concatenating the two hidden states from translation and language model for further processing. While such an approach can achieve certain improvement, it overlooks the potential of taking advantage of monolingual data into enhancing NMT training, since it is only used to obtain a language model.

Other studies attempt to enlarge the parallel bilingual training dataset through translating the monolingual data with a model trained by the given parallel corpora. Such an idea has been used both in NMT [7] and statistical machine translation [8]–[10]. Although this approach can increase the volume of parallel training data, it may introduce low-quality pseudo sentence pairs into the NMT training in the mean time, thus likely to hurt the performance of NMT model.

The concept of dual learning [11] was proposed to enhance the performance of translation models, in which two translation models teach each other through a reinforcement learning process by minimizing the reconstruction error of a monolingual sentence in either source or target languages. One potential issue of their approach is that it requires to back-propagate through the sequence of discrete predictions using reinforcement learning based approaches which are notoriously inefficient. Adopting the same idea of reconstruction error minimization, a reconstruction term was proposed to be appended to the training objective [12]. To some extent, the reconstruction methods could be seen as an iteration extension of pseudo sentence pair generating method, since after updating model parameters on the pseudo parallel corpus, the learned models are used to produce a better pseudo corpus.

In this work, motivated by the law of total probability, we propose a principled way to exploit monolingual data for NMT. According to the law of total probability, the marginal probability $P(y)$ can be computed using the conditional probability $P(y|x)$ in the following way: $P(y) = \sum_{x \in \mathcal{X}} P(y|x)P(x)$. As a result, ideally the learned conditional probability $P_\theta(y|x)$

Manuscript received October 29, 2018; revised May 12, 2019; accepted May 24, 2019. Date of publication June 6, 2019; date of current version July 12, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000904 and in part by the National Natural Science Foundation of China under Grants 61727809 and U1605251. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Taro Watanabe. *(Corresponding author: Enhong Chen.)*

Y. Wang and E. Chen are with the Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China, Hefei 230052, China (e-mail: wyjun@mail.ustc.edu.cn; cheneh@ustc.edu.cn).

Y. Xia, L. Zhao, J. Bian, T. Qin, and T.-Y. Liu are with the Microsoft Research Asia, Beijing 100091, China (e-mail: yingce.xia@gmail.com; lizo@microsoft.com; jiabia@microsoft.com; taoqin@microsoft.com; tyliu@microsoft.com).

Digital Object Identifier 10.1109/TASLP.2019.2921423

parameterized by θ should satisfy the following equation for any sentence y in target language:

$$P(y) = \sum_{x \in \mathcal{X}} P_\theta(y|x)P(x). \tag{1}$$

However, if $P_\theta(y|x)$ is learned using bilingual corpus via maximum likelihood estimation, there is no guarantee that the above equation will hold on monolingual corpus.

Therefore, assuming that both the parallel corpus and the monolingual corpus are sampled from the source and target language spaces $\mathcal{X}$ and $\mathcal{Y}$, in our previous work [13], we proposed to learn the translation model P_θ by maximizing the likelihood of parallel corpus subject to the constraint of Eqn.(1), for any target-language sentence y in a monolingual corpus $\mathcal{M}$. In this way, the learning objective can explicitly emphasize the probabilistic connection so as to regularize the learning process towards the right direction.

Further, when we pre-train the translation model from bilingual corpus, we could obtain a relatively well-trained model. Therefore, there is just a small gap between the two terms $P(y)$ and $\sum_{x \in \mathcal{X}} P_\theta(y|x)P(x)$ in equation Eqn.(1). So we use the term $\sum_{x \in \mathcal{X}} P_\theta(y|x)P(x)$ in the law of total probability as an estimation of the marginal distribution $P(y)$ and propose an alternative training objective as maximizing the likelihood of bilingual data and monolingual data simultaneously.

To compute $\sum_{x \in \mathcal{X}} P_\theta(y|x)P(x)$ in both training objectives, a technical challenge is that the value of this term is usually intractable due to the exponentially large search space $\mathcal{X}$. Traditionally, this problem can be resolved by sampling the full search space and using the sampled average to approximate the expectation:

$$\sum_{x \in \mathcal{X}} P_\theta(y|x)P(x) = \mathbb{E}_{x \sim P(x)} P_\theta(y|x)$$

$$\approx \frac{1}{K} \sum_{i=1}^{K} P_\theta(y|x^{(i)}), x^{(i)} \sim P(x). \tag{2}$$

That is, given a target-language sentence $y \in \mathcal{Y}$, one samples K source sentences $x^{(i)}$ according to distribution $P(x)$, and then computes the average conditional probability over the K samples. However, since the values of $P_\theta(y|x)$ are very sparse and most x from distribution $P(x)$ would get a nearly zero value for $P_\theta(y|x)$, a plain Monte Carlo sample from the distribution $P(x)$ may not be capable of regularizing the training of NMT models. To deal with this problem, we propose to sample from the distribution $P(x|y)$ instead of $P(x)$, and adopt the method of *importance sampling* to guarantee the quality of sampled sentences such that the corresponding constraint is valid empirically. The training process is illustrated in Figure 1, where $\text{NMT}_{x \to y}$ and $\text{NMT}_{y \to x}$ denote translation models $P_\theta(y|x)$ and $P(x|y)$ respectively.

The main contributions of this paper can be summarized as follows:

- We propose a new semi-supervised approach for NMT, which adopts a probabilistic view using the law of total probability to leverage monolingual data to enhance the training of NMT.
- When estimating the expectation term $\sum_{x \in \mathcal{X}} P_\theta(y|x)P(x)$ in the law of total probability, we adopt importance sampling to guarantee the quality of generated sentences and ensure that the probabilistic constraint is valid empirically.
- Experiments on the IWSLT and WMT datasets show that our approach can achieve significant improvement in terms of translation quality over existing semi-supervised NMT approaches on both German→English and English→French translation tasks.

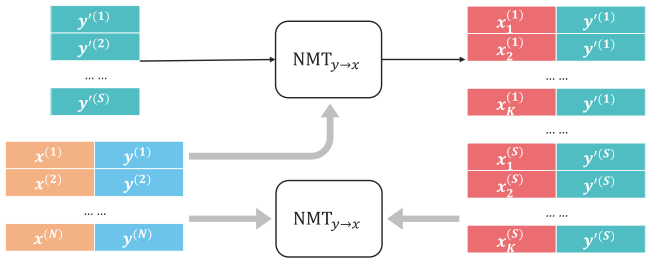


Fig. 1. Illustration of the proposed method. Specifically, the left upper y denotes monolingual data, the left bottom x and y denote bilingual data, and the x and y in the right are sampled data from monolingual data by $\text{NMT}_{y \to x}$.

II. Related Work

Neural machine translation has drawn much attention in recent years. For the standard NMT system, only bilingual corpora are used for model training with MLE method. Since it is costly to collect bilingual data, exploring monolingual data for machine translation has attracted intensive attention. The methods proposed for this purpose could be divided into three categories: (1) integrating language model trained with monolingual data into NMT model, (2) generating pseudo sentence pairs from monolingual data and (3) jointly training of both source-to-target and target-to-source translation models by minimizing reconstruction errors of monolingual sentences.

In the first category, a language model separately trained with monolingual data is integrated into the NMT model. Language model in the target language can promote the ability of NMT model mainly because that it could increase the score of fluent outputs during decoding of NMT model [14]. Specifically, language models were trained independently with target-side monolingual sentences, and incorporated into the neural network during decoding by rescoring of the beam or adding the recurrent hidden state of the language model to the decoder states [4]. Different from these methods, Cold Fusion method [15] encourages the Seq2Seq decoder to learn to use the external language model during training, and shows its effectiveness on the speech recognition task. Neural architecture also allows multi-task learning and parameter sharing between MT and target-side LM [16], and multi-task learning has shown to be effective in the context of sequence-to-sequence models where different parts of the network can be shared across multiple tasks [17]. Language models trained on monolingual corpora were also used to initialize both encoder and decoder networks of NMT model [18].

In the second category, monolingual data is translated using translation model trained from bilingual sentence pairs, and being paired with its translations to form a pseudo parallel corpus to enlarge the training data. Specifically, several authors have explored back-translating target-side monolingual data to produce synthetic parallel data for phrase-based SMT [8], [9]. Similar approach also has been applied to NMT, and back-translated synthetic parallel data has been found to have a more general use in NMT than in SMT, with positive effects that go beyond domain adaption [7]. Further, the understanding of back-translation was broadened with large scale training corpora and a number of methods to generate synthetic source sentences [19]. Recently, a multi-task learning framework was proposed to exploit source-side monolingual data, in which machine translation on synthetic bilingual data and sentence reordering with source-side monolingual data were jointly performed [20]. Moreover, both source and target monolingual data was explored for reinforcement learning training [21]. For these methods, due to the imperfection of machine translation system, some of the incorrect translations are very likely to hurt the performance of source-to-target model [11], [22]. As an extension of back-translation methods, translation probabilities from target-to-source model were introduced as weights of synthetic parallel sentences to punish poor pseudo parallel sentences, and further interactive training of NMT models in two directions were used to refine them [22]. The negative impact of noisy translations can be minimized since the generated synthetic sentence pairs are weighted.

In the third category, monolingual data is reconstructed with both source-to-target and target-to-source translation models, and then the two models are jointly trained. Specifically, a reconstruction term was appended to the training objective, which aims to reconstruct the observed monolingual corpora using an autoencoder [12]. In another work [11], two translation models taught each other through a reinforcement learning process, based on the feedback signals generated during this process. To some extent, the reconstruction methods could be seen as an iteration extension of [7]'s method, since after updating model parameters on the pseudo parallel corpus, the learned models are used to produce a better pseudo corpus.

III. Background: Neural Machine Translation

Neural machine translation systems are typically implemented based on an encoder-decoder neural network framework, which learns a conditional probability $P(y|x)$ from a source language sentence x in space $\mathcal{X}$ to a target language sentence y in space $\mathcal{Y}$. In this framework, the encoder neural network projects the source sentence into a distributed representation, based on which the decoder generates the target sentence word by word. The encoder and the decoder are learned jointly in an end-to-end way. The standard training objective of existing NMT models is to maximize the likelihood of the training data.

With fast development of deep learning, a variety of encoder-decoder architectures have been introduced to enhance the NMT performance, such as recurrent neural networks (RNN) with attention mechanisms [1], [23], [24], convolutional neural network (CNN) based frameworks [3], [25], and, most recently, transformer framework [26]. In the mean time, a trend of recent works is to focus on improving NMT by increasing the model depth, since deeper neural networks usually imply stronger modeling capability [2], [27]. However, even a single layer NMT model has a huge number of parameters to optimize, which requires large-scale data for effective model training, not to mention deep models. Unfortunately, parallel bilingual corpora are usually quite limited in either quantity or coverage, making it appealing to exploit large-scale monolingual corpora to improve NMT.

IV. Framework

In this section, we present our semi-supervised approach for training NMT models which leverages monolingual data from a probabilistic perspective. We first introduce the inherent relationship between monolingual data and the NMT model brought by the law of total probability, and propose two training objectives leveraging the monolingual data. Given the difficulty in estimating the expectation term in the law of total probability brought by the exponentially large search space, we then propose to address this challenge by the technique of importance sampling. After that, we present a whole semi-supervised algorithm for NMT in detail.

A. Training Objective

We introduce two new training objectives, both of which leverage the law of total probability: one is to minimize the gap between two estimations of the probabilities, and the other is to directly maximize the marginal probabilities of monolingual data. Note that in the above two cases, the NMT model is involved to calculate the corresponding probabilities.

Objective 1: Given the source language space $\mathcal{X}$ and target language space $\mathcal{Y}$, a translation model takes a sample from $\mathcal{X}$ as input and maps to space $\mathcal{Y}$. In common practice, the translation model is represented by a conditional distribution $P_\theta(y|x)$ parameterized by θ, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In standard supervised learning, given a parallel corpus with N sentence pairs $\mathcal{B} = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$, the translation model is learned by maximizing the likelihood of the training data:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log P_\theta(y^{(n)}|x^{(n)}). \tag{3}$$

On the other hand, given the law of total probability $P(y) = \sum_{x\in\mathcal{X}} P(y|x)P(x)$, for any $y \in \mathcal{Y}$, if the learned translation model θ is perfect, we should have:

$$P(y) = \sum_{x\in\mathcal{X}} P_\theta(y|x)P(x) = \mathbb{E}_{x\sim P(x)} P_\theta(y|x), \tag{4}$$

which connects sentence y to the translation model $P_\theta(y|x)$.

Assume that we have a monolingual corpus $\mathcal{M}$ which contains S sentences i.i.d. sampled from the space $\mathcal{Y}$ according to marginal distribution $P(y')$, i.e., $\mathcal{M} = \{y'^{(s)}\}_{s=1}^{S}$ where $y' \in \mathcal{Y}$. Considering the model P_θ is empirically learned via MLE training from parallel data, there is no guarantee that Eqn.(4) will hold for sentences in $\mathcal{M}$. Therefore, we can regularize the learning

process on monolingual data by forcing all sentences in $\mathcal{M}$ to satisfy the probabilistic relation in Eqn.(4), which guides the model to a better direction. Mathematically, we can formulate the aforementioned training mechanism as the following constrained optimization problem:

$$\max \sum_{n=1}^{N} \log P_\theta(y^{(n)}|x^{(n)}),$$
$$\text{s.t.} \quad P(y') = \mathbb{E}_{x\sim P(x)} P_\theta(y'|x), \forall y' \in \mathcal{M}. \tag{5}$$

Then, we propose our training objective according to the proposed constrained optimization problem in Eqn.(5). Following the common practice in constrained optimization, we convert the constraint into the following regularization term:

$$S_1(\theta) = [\log P(y') - \log \mathbb{E}_{x\sim P(x)} P_\theta(y'|x)]^2, \tag{6}$$

and then add it to the maximum likelihood training objective. Formally, we introduce our new semi-supervised training objective as minimizing the following loss function:

$$\mathcal{L}_{S_1}(\theta) = -\sum_{n=1}^{N} \log P_\theta(y^{(n)}|x^{(n)}) + \lambda_1 \sum_{s=1}^{S} [\log P(y'^{(s)}) - \log \mathbb{E}_{x\sim P(x)} P_\theta(y'^{(s)}|x)]^2, \tag{7}$$

where λ_1 is the hyper-parameter controlling the tradeoff between the likelihood and regularization term.

Objective 2: Another commonly adopted way to deal with monolingual data samples in machine learning literature is to maximize their probabilities [28], [29], i.e.,

$$\mathcal{L}_{S_2}(\theta) = -\sum_{n=1}^{N} \log P_\theta(y^{(n)}|x^{(n)}) - \lambda_2 \sum_{s=1}^{S} \log P(y'^{(s)}), \tag{8}$$

where λ_2 is the hyper-parameter controlling the tradeoff between the likelihood of bilingual data and the likelihood of monolingual data.

Leveraging the law of total probability, we can incorporate the NMT model into the above equation by the following way:

$$\mathcal{L}_{S_2}(\theta) = -\sum_{n=1}^{N} \log P_\theta(y^{(n)}|x^{(n)}) - \lambda_2 \sum_{s=1}^{S} \log \mathbb{E}_{x\sim P(x)} P_\theta(y'^{(s)}|x). \tag{9}$$

Empirical Adaption: Since the ground-truth marginal distributions $P(x)$ and $P(y)$ are usually not available, we use the empirical marginal distributions $\hat{P}(x)$ and $\hat{P}(y)$ as their proxies, which we get from well-trained language models. Then, the proposed training objectives become:

$$\mathcal{L}_{S_1}(\theta) = -\sum_{n=1}^{N} \log P_\theta(y^{(n)}|x^{(n)}) + \lambda_1 \sum_{s=1}^{S} [\log \hat{P}(y'^{(s)}) - \log \mathbb{E}_{x\sim \hat{P}(x)} P_\theta(y'^{(s)}|x)]^2, \tag{10}$$

$$\mathcal{L}_{S_2}(\theta) = -\sum_{n=1}^{N} \log P_\theta(y^{(n)}|x^{(n)}) - \lambda_2 \sum_{s=1}^{S} \log \mathbb{E}_{x\sim \hat{P}(x)} P_\theta(y'^{(s)}|x). \tag{11}$$

B. Estimation via Importance Sampling

We can see that for both training objectives, we need to compute the expectation term $\mathbb{E}_{x\sim \hat{P}(x)} P_\theta(y'|x)$ for each sentence y' in the monolingual corpus. To compute this term, a technical challenge arises as this expectation is usually intractable due to the exponential search space of x. A straightforward way to address such large search space problem is to build an approximate estimator by sampling the full search space via Monte Carlo technique. That is, if we sample K sentences from distribution $\hat{P}(x)$, an empirical estimation of $\mathbb{E}_{x\sim \hat{P}(x)} P_\theta(y'|x)$ can be computed as $\frac{1}{K}\sum_{i=1}^{K} P_\theta(y'|x^i)$.

However, there exists a problem when we estimate the expectation term by sampling from distribution $\hat{P}(x)$. Intuitively, given a certain sentence y' in the target language, when we sample a sentence x in the source language from empirical marginal distribution $\hat{P}(x)$ through conforming a good source side language model, it is almost impossible that x is exactly or close to the translation of y'. In other words, the sampled sentence x from empirical marginal distribution $\hat{P}(x)$ in source language is usually irrelevant to a certain sentence y' in target language. Formally, since $P_\theta(y'|x)$ yields a severe uneven distribution over space $\mathcal{X}$, for a pre-trained translation model P_θ, most of those samples from distribution $\hat{P}(x)$ would result in $P_\theta(y'|x)$ very close to zero. This will make the constraint empirically invalid to regularize a better model P_θ, since actually we want to train a translation model which could better model the conditional distribution $P(y|x)$ when the source sentence x and target sentence y are the translation for each other, while we just don't care about the conditional distribution $P(y|x)$ when x and y are irrelevant. Therefore, in order to make the constraint effective, we should get samples that can achieve relatively large $P(y|x)$, i.e., making sampled sentences x relevant to the given sentence y. Similarly, for the objective of maximizing both the likelihood of monolingual data and bilingual data, we also need to make sampled sentence x relevant to the given sentence y' in order to obtain a better translation model P_θ.

Given a target sentence y, to get sentences x relevant to y, intuitively we could obtain a translation of sentence y using a pre-trained translation model which maps input sentence y in space $\mathcal{Y}$ to a sentence in space $\mathcal{X}$. Therefore, we propose to get

related source sentence x by generating from a reverse direction translation model $P(x|y)$. In this way, we can get constraint on $P_\theta(y'|x)$ with large probability, making our constraint valid empirically. Since we sample from distribution $P(x|y)$ instead of $\hat{P}(x)$ when estimating $\mathbb{E}_{x\sim\hat{P}(x)}P_\theta(y'|x)$, we have to adjust our estimate somehow to account for having sampled from the distribution $P(x|y)$. Specifically, we can rewrite $\mathbb{E}_{x\sim\hat{P}(x)}P_\theta(y'|x)$ as follows:

$$\begin{aligned}\mathbb{E}_{x\sim\hat{P}(x)}P_\theta(y'|x) &= \sum_{x\in\mathcal{X}} P_\theta(y'|x)\hat{P}(x) \\ &= \sum_{x\in\mathcal{X}} \frac{P_\theta(y'|x)\hat{P}(x)}{P(x|y')}P(x|y') \\ &= \mathbb{E}_{x\sim P(x|y')}\frac{P_\theta(y'|x)\hat{P}(x)}{P(x|y')}. \end{aligned} \tag{12}$$

That is, by making a multiplicative adjustment to $P_\theta(y'|x)$ we compensate for sampling from $P(x|y)$ instead of $\hat{P}(x)$. This procedure is exactly the technique of *importance sampling* [30]–[32]. Then, the *importance sampling estimation* of $\mathbb{E}_{x\sim\hat{P}(x)}P_\theta(y'|x)$ is

$$\frac{1}{K}\sum_{i=1}^{K}\frac{P_\theta(y'|x_i)\hat{P}(x_i)}{P(x_i|y')}, x_i \sim P(x|y') \tag{13}$$

where K is the sample size. Based on Eqn.(12), we can use any sampling approach to estimate the expectation. Considering that random sampling brings very large variance and sometimes unreasonable results in machine translation, we use beam search to obtain more meaningful results as in [11], [22].

Therefore, empirically our semi-supervised training objectives are:

$$\begin{aligned}\mathcal{L}_{E_1}(\theta) = & -\sum_{n=1}^{N}\log P_\theta(y^{(n)}|x^{(n)}) \\ & + \lambda_1\sum_{s=1}^{S}\Big[\log\hat{P}(y'^{(s)}) \\ & - \log\frac{1}{K}\sum_{i=1}^{K}\frac{\hat{P}(x_i^{(s)})P_\theta(y'^{(s)}|x_i^{(s)})}{P(x_i^{(s)}|y'^{(s)})}\Big]^2, \end{aligned} \tag{14}$$

$$\begin{aligned}\mathcal{L}_{E_2}(\theta) = & -\sum_{n=1}^{N}\log P_\theta(y^{(n)}|x^{(n)}) \\ & - \lambda_2\sum_{s=1}^{S}\log\frac{1}{K}\sum_{i=1}^{K}\frac{\hat{P}(x_i^{(s)})P_\theta(y'^{(s)}|x_i^{(s)})}{P(x_i^{(s)}|y'^{(s)})}. \end{aligned} \tag{15}$$

C. Algorithm

We learn the model $P_\theta(y|x)$ by minimizing the training objectives in Eqn.(14) and Eqn.(15), i.e., the weighted combination between the likelihood of bilingual data and the marginal distribution regularization term as shown in Eqn.(14), and the combination between the likelihood of bilingual data and monolingual data as shown in Eqn.(15), respectively. The details of our proposed algorithm is shown in Algorithm 1. The input of this algorithm consists of a monolingual corpus $\mathcal{M}$ containing sentences in the target language B, a bilingual corpus containing sentence pairs in language A and language B, empirical marginal distributions $\hat{P}(x)$ and $\hat{P}(y)$, and an existing translation model that can translate sentences in language B to language A. The translation model we want to learn and the reverse direction translation model used for sampling are denoted as $P_\theta(y|x)$ and $P(x|y)$ respectively. Before training procedure begins, we sample K sentences according to the translation model $P(x|y)$ for each sentence in the monolingual corpus. During training, in one mini-batch, we get m sentences from $\mathcal{M}$ and b sentence pairs from $\mathcal{B}$. Then, we compute the gradient of the objective function with to the parameter θ and finally update the parameter θ.

Algorithm 1: Semi-supervised NMT via Marginal Distribution Estimation.

Input: Monolingual corpus $\mathcal{M}$, bilingual corpus $\mathcal{B}$, translation model $P(x|y)$, empirical marginal distributions $\hat{P}(x)$ and $\hat{P}(y)$, hyper-parameters λ_1 or λ_2, sample size K.

1: Initialize translation model $P_\theta(y|x)$ with random weights θ.
2: Pre-train translation model $P_\theta(y|x)$ by maximizing $\log P_\theta(y|x)$ on bilingual corpus $\mathcal{B}$.
3: For each sentence y' in $\mathcal{M}$, generate K sentences $\hat{x}_1, ...\hat{x}_K$ according to the translation model $P(x|y)$;
4: **repeat**
5: Get a mini-batch of monolingual sentences M from $\mathcal{M}$ where $|M| = m$, and a mini-batch of bilingual sentence pairs B_{AB} from $\mathcal{B}$ where $|B_{AB}| = b$;
6: Calculate the semi-supervised training objectives $\mathcal{L}_{E_1}$ or $\mathcal{L}_{E_2}$ according to Eqn.(14) or Eqn.(15) based on B_{AB}, M and the corresponding translations;
7: Update the parameters of θ:

$$\theta \leftarrow \theta - \gamma\nabla_\theta\mathcal{L}_{E_1}(\theta) \tag{16}$$

or

$$\theta \leftarrow \theta - \gamma\nabla_\theta\mathcal{L}_{E_2}(\theta) \tag{17}$$

8: **until** model converged

D. Discussion

We first analyze the relationship of the proposed two objectives. Specifically, the marginal distribution regularization objective assumes that the law of total probability wouldn't hold for sentences in monolingual corpus, and enforces it to be guaranteed via a regularization term, while the likelihood maximization objective assumes that after pre-training on bilingual data, the law of total probability approximately holds for sentences in monolingual data. It may sound contradictory on the two training objectives, while in fact, after pre-training, we could obtain a relatively good translation model, and there is just a small gap

between the two terms in the law of total probability. In this situation, both enforcing the equality to hold and maximizing the estimated likelihood of monolingual sentences will lead the translation model to a better direction. The empirical analysis in the training procedure for both training objectives will be shown in the experiment section, and the experiment results support our assumption.

Note that when training the source-to-target translation model, we use a pre-trained target-to-source translation model to generate samples for a target sentence y in monolingual corpus. This procedure is similar to back-translating for enlarging the parallel bilingual training dataset through translating the monolingual data [7]–[10]. Our proposed method is different from these works since they directly use the generated pseudo sentence pairs as a supplement to bilingual data and train the translation model via maximum likelihood estimation, while our method uses the generated sentences to compute the probability of each component in the law of total probability. We believe that this difference could rescue our method from suffering from the problem of low-quality pseudo sentence pairs which may limit the performance of back-translation methods. This is because that the back-translation methods treat generated pseudo sentence pairs the same as bilingual data, causing low-quality pseudo sentence pairs leading the translation model to a wrong direction. On the other hand, since the law of total probability used in our method is an inherent property for any sentence pair, the quality of the generated sentence will not have much effective on the translation model.

Further, a class of works aiming to reconstruct the monolingual data with both source-to-target and target-to-source translation models and jointly train the two models also use the reverse direction model to generate samples [11], [12]. In these reconstruction-based methods, it is supposed that the two models can get mutual benefit from this iterative process, while in practice this process is not easy to control since it may suffer from an error propagation problem. Mistakes made in source-to-target translation will be propagated to target-to-source translation. Compared with reconstruction-based approaches, our method achieves direct modeling of the NMT model by exploiting the property of probability. Since our method just uses the reverse direction model to sample one time, it will not suffer from the error propagation problem, and thus, the learning process is more stable. One may argue that our approach could be used to improve the reversed translation models as well. Actually, when improving a translation model, our approach uses samples from an arbitrary reversed model. Therefore, we don't need to iteratively sample from both models, and we wouldn't suffer from the error propagation problem. However, if we do alternate the optimizations by iteratively sampling from both models, we may suffer from the similar problems.

Our proposed marginal distribution regularization objective in Eqn.(14) leverages the relationship between translation models and language models in each language space by exploiting the law of total probability. This is similar to the work [33] which uses the probability property as a regularization term in supervised tasks which are emerged in dual forms. Specifically, they proposed training the models of two dual tasks simultaneously, and explicitly exploiting the probabilistic correlation between them to regularize the training process using the following equality:

$$P(x)P(y|x) = P(y)P(x|y). \tag{18}$$

Although our method explicitly exploits the probabilistic correlation as well, it is different from their work since we guarantee the law of total probability on monolingual data, while they exploited the correlation on bilingual data.

Similar to our likelihood maximization objective in Eqn.(15), [22] also maximized the likelihood of both bilingual data and monolingual data. However, they treated the source translations as hidden states for the target sentences, and optimized a lower bound of the true likelihood function with EM algorithm, while we computed the true likelihood function using the law of total probability. This means that we can obtain a more accurate estimation of the likelihood of monolingual data than their method.

Considering the longer training time and larger memory consumption of our method compared to MLE training, actually, to leverage monolingual data in the training procedure of NMT, existing semi-supervised NMT methods could be roughly divided into two categories: language model combining and data augmentation methods. Specifically, language model combining method trains separate language models with monolingual data, then integrates the trained language models into the NMT model (by rescoring of the beam or adding the recurrent hidden state of the language model to the decoder states). Compared to MLE, this kind of method doesn't need extra training time and memory consumption except for training language models. However, the performance of language model combining method is quite limited since it doesn't fundamentally address the shortage of parallel training data. As for data augmentation methods, generally they adopt different training objectives or strategies with the same procedure of data augmentation (i.e., translating source side monolingual data with source-target translation model or translating target side monolingual data with target-source translation model.) This category of methods need longer training time and larger memory consumption mainly because of data augmentation, and training time or memory consumption mainly depends on the critical hyper-parameter sample size (i.e., the number of translated sentences for a given monolingual sentence). Fortunately, compared to iteratively training of NMT models, our method only needs to translate target side monolingual data once using pre-trained target-source translation model, thus saving much time. To sum up, to effectively leverage monolingual data in semi-supervised NMT, data augmentation is crucial in the literature, thus longer training time and larger memory consumption are trade-off for performance improvement. Further, our method can save much training time compared to iteratively training methods.

V. Experiments

We conducted a set of experiments on two translation tasks to demonstrate the performance of the proposed methods.

A. Settings

Datasets: We evaluated our approach on two translation tasks: English→French (En→Fr) and German→English (De→En). Specifically, for En→Fr task, we used a subset of the bilingual corpus from WMT'14 for training, which contains 12M sentence pairs extracted from five datasets: Europarl v7, Common Crawl corpus, UN corpus, News Commentary, and 10^9 French-English corpus. Following common practices, we concatenated newstest2012 and newstest2013 as the validation set, and used newstest2014 as the test set. The validation and test sets for En→Fr contain 6k and 3k sentence pairs respectively. We used the "News Crawl: articles from 2012" provided by WMT'14 as monolingual data. We used 5M monolingual sentences to train our model. For De→En task, the bilingual corpus is from IWSLT 2014 evaluation campaign [34], as used in [35] and [36], containing about 153k sentence pairs for training, and 7k/6.5k sentence pairs for validation/test. The monolingual data for De→En is the latest version of the TED talks corpus, which is available on the WIT^3 webset [37]. The amount of monolingual sentences for De→En is about 150k after preprocessing. During the training of both tasks, we drop all sentences with more than 50 words.

Empirical Marginal Distribution $\hat{P}(x)$ and $\hat{P}(y)$: We used LSTM-based language modeling [5], [38] approach to characterize the marginal distribution of a given sentence. Specifically, for En→Fr, we used a single layer LSTM with word embeddings of 512 dimensions and hidden states of 1024 dimensions respectively. For De→En, we trained a language model with 512 dimensions for both word embeddings and hidden states. The batch sizes for En→Fr and De→En were 128 and 256 respectively during training. The vocabularies (including whether subword units were used) for each language model were the same as those for corresponding translation tasks. Both the models in the two translation tasks were trained using Adam [39] as [40] with initial learning rate 0.0002. The language models were fixed during the training procedure of translation models.

Implementation Details: For En→Fr translation, we implemented a basic single-layer RNNSearch model [1], [41] to ensure fair comparison with the related works [1], [7], [11], and a deep LSTM model to see improvement brought by our algorithm combining with more recent techniques. Note that "deep" is in comparison with single layer structures. Specifically, for the basic RNNSearch model, we followed the same setting as that in related works following the common practice. To be more specific, GRUs were applied as the recurrent units. The dimensions of word embedding and hidden state were 620 and 1000 respectively. We constructed the vocabulary with the most common 30K words in the parallel corpora. Out-of-vocabulary words were replaced with a special token $\langle$UNK$\rangle$. For monolingual corpora, we removed the sentences containing out-of-vocabulary words. In order to prevent over-fitting, we applied dropout during training [42], where the dropout probability is 0.1. We leveraged an open source NMT system implemented by Theano[1] for the experimnets. For the deep LSTM model, the dimensions of embedding and hidden states are 512 and 1024 respectively. Both the encoder and decoder have four stacked layers with residual connections [43]. To deal with out-of-vocabulary problem, we adopted the byte-pair encoding (BPE) techniques [44] to split words into sub-words with 32000 BPE operations, which can efficiently address rare words.[2]

For De→En translation, we implemented a two-layer LSTM model with both word embedding and hidden state dimensions 256. We applied dropout with probability 0.1. We also adopted BPE to split the words with 25000 BPE operations.

The estimation of parameter number for the single-layer RNNSearch, 4-layer LSTM and 2-layer LSTM we used are 102M, 130M and 46M respectively. It is worth mentioning that our method is capable for any NMT model structures, and the number of parameters depends on which model structure we adopt. Further, to fairly compare our method with other semi-supervised NMT methods, we used the same structure for different methods in each translation task, so the number of parameters of the translation model for each method is just the same.

Note that each task needs a reverse translation model. We trained a Fr→En NMT model with test BLEU 35.46 and a En→De model with test BLEU 23.94.

Baseline Methods: We compared our approach with several strong baselines, including a well known attention-based NMT system *RNNSearch* [1], a deep LSTM structure, and several semi-supervised NMT models:

- *RNNSearch:* For En→Fr translation, we exactly followed the settings reported in [1]. Only bilingual corpora were used to train a standard attention-based NMT model. The obtained RNNSearch model was used as initialization for semi-supervised models.
- *deep LSTM:* We trained a four-layer LSTM model for En→Fr translation and a two-layer LSTM model for De→En translation respectively. Only bilingual corpora are used during training. The obtained LSTM models were also used to initialize semi-supervised algorithms.
- *shallow fusion-NMT:* This method incorporates a target-side language model which is trained using monolingual corpora into the translation model during decoding by rescoring the beam, named as shallow fusion [4].
- *pseudo-NMT:* Bilingual and target-side monolingual corpora were used. This method generates pseudo bilingual sentence pairs from monolingual corpora to assist training [7]. We used the same reverse NMT model to generate pseudo bilingual sentence pairs as the sampling model in our method.
- *dual-NMT:* Bilingual and target-side monolingual corpora were used. This method reconstructs the monolingual data with both source-to-target and target-to-source translation models and jointly trains the two models with dual learning objective [11].

Training Procedure: Following [11], [45], to speed up training, for each task, we first trained NMT models on their corresponding parallel corpora and then ran our algorithm with the obtained models as initialization.

To obtain the models used to initialize our algorithm, (1) for the single-layer RNNSearch model in En→Fr translation, we

[1]https://github.com/nyu-dl/dl4mt-tutorial

[2]https://github.com/rsennrich/subword-nmt

TABLE I
TRANSLATION RESULTS OF EN→FR AND DE→EN TRANSLATION TASKS. THE NMT SETTING REPRESENTS STANDARD MLE TRAINING OBJECTIVE WITH ONLY BILINGUAL DATA. THE MODEL STRUCTURE FOR EN→FR IS THE RNNSEARCH MODEL [1]. THE MODEL STRUCTURE FOR DE→EN IS THE TWO-LAYER LSTM MODEL. THE INITIALIZATION MODELS FOR ALL SEMI-SUPERVISED NMT SYSTEMS ARE CONSISTENT FOR EACH LANGUAGE PAIR, I.E., RNNSEARCH FOR EN→FR TRANSLATION AND TWO-LAYER LSTM FOR DE→EN TRANSLATION RESPECTIVELY. Δ MEANS THE IMPROVEMENT OVER STANDARD NMT

System	English→French	Δ	German→English	Δ
NMT	29.92		30.99	
Representative semi-supervised NMT systems				
shallow fusion-NMT [4]	30.03	+0.11	31.08	+0.09
pseudo-NMT [7]	30.40	+0.48	31.76	+0.77
dual-NMT [11]	32.06	+2.14	32.05	+1.06
Our semi-supervised NMT systems				
objective 1: marginal distribution regularization	**32.85**	**+2.93**	**32.35**	**+1.36**
objective 2: monolingual likelihood estimation	**32.82**	**+2.90**	**32.24**	**+1.25**
objective 1 + objective 2	**32.89**	**+2.97**	**32.41**	**+1.42**

followed the same training procedure as that proposed by [46]; (2) for deep LSTM architectures, we trained the model with mini-batch size 128 for En→Fr translation and 32 for De→En translation respectively. Gradient clipping was used with clipping value 1.0 and 2.5 for En→Fr and De→En respectively [47]. Models were optimized by AdaDelta [48] on M40 GPU until convergence.

To run our algorithm, for both training objectives, we used AdaDelta with the mini-batch of 32 bilingual sentence pairs and 32 monolingual sentences for all tasks. The sample size K, the hyper-parameter λ_1 and the hyper-parameter λ_2 in our method were set as 2, 0.05 and 2 respectively according to the trade-off between validation performance and training time for all tasks. Further, in addition to training with the two proposed objectives separately, we propose to combine the two objectives to enhance translation performance. Specifically, we first run our algorithm with the marginal distribution objective until the performance stopped to improve on the validation set. Then, we continued to run our algorithm with the second semi-supervised training objective to see if the performance has improvement.

To be fair in all experiments, for each translation task, *pseudo-NMT* and *dual-NMT* adopted the same settings as our approach including the same source and target monolingual corpora, and the initialization models.

Evaluation Metrics: The translation qualities are measured by case-insensitive BLEU [49] as calculated by the *multi-bleu.perl* script,[3] which is widely used in machine translation and other tasks [50]. A larger BLEU score indicates a better translation quality. During testing, for the single-layer model in En→Fr translation, following the common practice, we used beam search [51] with beam size 12 as in many previous works; for deep LSTM structures in both En→Fr and De→En translation, the beam size was set to 5.

B. Main Results

We report the experiment results in this subsection.

Table I shows the results of our method and three semi-supervised baselines with the aligned network structures. We can see that our method of both training objectives outperforms baseline algorithms on both language pairs. For the translation from English to French, our marginal distribution regularization objective outperforms the RNNSearch model with MLE training objective by 2.93 points, and outperforms the strongest baseline dual-NMT among all single layer models by 0.79 point. On the other hand, our objective of maximizing the likelihood of both monolingual and bilingual data outperforms the standard RNNSearch model and dual-NMT by 2.90 and 0.76 points respectively. Further, the combination of the proposed two training objectives outperforms the RNNSearch model by 2.97 points. For the translation from German to English, we got similar results. Specifically, our method with two training objectives outperforms RNNSearch with MLE training objective by 1.36 and 1.25 points, and outperforms dual-NMT by 0.3 and 0.19 points, respectively. Moreover, we can see that the combination of two objectives outperforms RNNSearch by 1.42 points. Improvements brought by our algorithm are significant compared with standard NMT. These results demonstrate the effectiveness of our algorithm.

Table II shows the comparison between our proposed algorithm incorporated in a deep model and several advanced baselines on En→Fr translation task. We can see that given a strong baseline, our algorithm can still make significant improvement, i.e., from 38.80 to 39.98 and 39.90 respectively with different training objectives. The combination of proposed two objectives achieves a BLEU score of 40.01. We also compared our method with the state-of-the-art back-translation baseline [19] for the large-scale WMT14 En→Fr scenario with 4-layer LSTM in Table II. Specifically, to better calibrate the effectiveness of our method, the experiments include: (1) back-translating two source sentences for each target monolingual sentence with beam search, which is the same as our method, (2) back-translating one source sentence for each target monolingual sentence with sampling, which is verified more effective than beam search in [19]. From Table II we can see that when back-translating two source sentences for each target monolingual sentence with beam search, our methods outperform the back-translation method. Moreover, our method also outperforms the superior sampling based back-translation method.

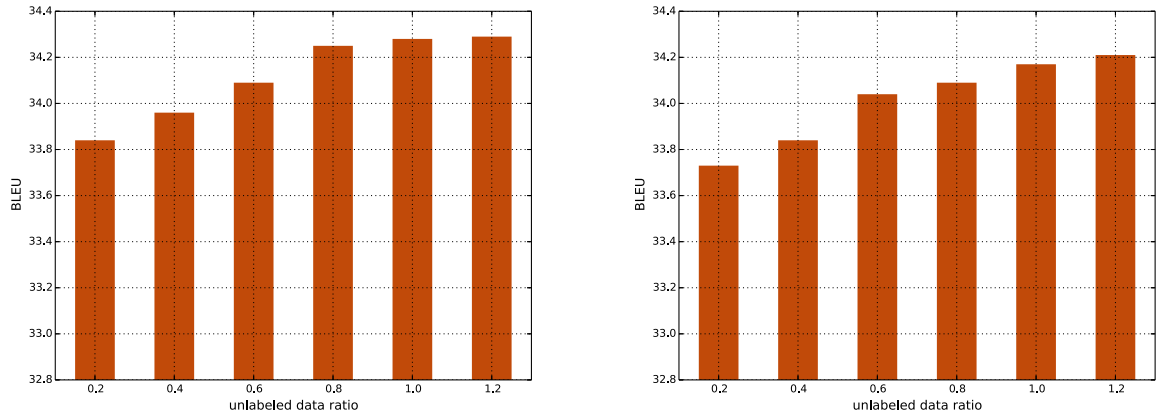
Furthermore, one may be curious about the question that given a parallel corpus, how many unlabeled sentences are most beneficial for improving translation quality? To answer this question, we investigated the impact of unlabeled data ratio on translation quality, which is defined as the number of unlabeled sentences

[3]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

TABLE II
DEEP NMT SYSTEMS' PERFORMANCES ON ENGLISH→FRENCH TRANSLATION

System	System Configuration	BLEU
	Representative end-to-end NMT systems	
[3]	15-15 layers CNN + BPE + 12M parallel data	38.45
[2]	8-8 layers LSTM (1024*1024 size) + BPE + 36M parallel data	38.95
[28]	9-7 layers LSTM + PosUNK +36M parallel data	39.2
[27]	Transformer (base model) + BPE + 36M parallel data	38.1
[27]	Transformer (big) + BPE + 36M parallel data	41.8
	Representative semi-supervised NMT systems	
[20]	Transformer (big) + BPE + 36M parallel data +31M monolingual data (sampling)	45.6
[20]	4-4 layers LSTM (512*1024 size) + BPE + 12M parallel data + 5M Monolingual Data (sampling)	39.51
[20]	4-4 layers LSTM (512*1024 size) + BPE + 12M parallel data + 5M Monolingual Data (beam search 2 sentences)	39.42
	Our deep NMT baseline	
this work	4-4 layers LSTM (512*1024 size) + BPE + 12M parallel data	**38.80**
	Our semi-supervised NMT systems	
this work	4-4 layers LSTM (512*1024 size) + BPE + 12M parallel data + 5M Monolingual Data (objective 1)	**39.98**
	4-4 layers LSTM (512*1024 size) + BPE + 12M parallel data + 5M Monolingual Data (objective 2)	**39.90**
	4-4 layers LSTM (512*1024 size) + BPE + 12M parallel data + 5M Monolingual Data (objective 1 + objective 2)	**40.01**

* The BLEU scores with underlines are reported in corresponding papers.



(a) Impact of unlabeled data ratio for marginal distribution regularization objective.

(b) Impact of unlabeled data ratio for the objective of maximizing the likelihood of both monolingual and bilingual data.

Fig. 2. Impact of unlabeled data ratio on De→En validation set.

divided by the number of labeled sentence pairs. Figure 2 shows the BLEU scores of the De→En validation set with different unlabeled data ratios. We constructed monolingual corpora with unlabeled data ratio from 0.2 to 1.2. We find that when unlabeled data ratio is no more than 0.8, increasing unlabeled data ratio leads to apparent promotion on translation quality, while the promotion tends to be inapparent when unlabeled data ratio exceeds 0.8. Therefore, to consider the balance between model performance and training efficiency, we didn't use monolingual data with unlabeled data ratio larger than 1.2, and leave it a future work to use more monolingual data.

For the time complexity of the training procedure, it is obvious that our approach needs less training time than dual-NMT, since we only train one model while dual-NMT works on two. Our approach takes almost the same training time as that of pseudo-NMT, since the terms $\hat{P}(y'^{(s)})$, $\hat{P}(x_i^{(s)})$ and $P(x_i^{(s)}|y'^{(s)})$ in Eqn. (14) and Eqn. (15) can be calculated offline. Then, during training, the time consumption of our method of one update depends on sample size K and batch size. Specifically, assuming that in one mini-batch, we get N_m sentences from monolingual corpus and N_b sentence pairs from bilingual corpus, then the computational complexity should be $O(N_b + KN_m)$, since the empirical marginal distribution $\hat{P}(y'^{(s)})$ and $\hat{P}(x_i^{(s)})$ as well as the probability $P(x_i^{(s)}|y'^{(s)})$ are calculated offline. For the pseudo-NMT method, if batch size is denoted as N, then the computational complexity is $O(N)$. Due to the constraint of GPU memory, N and $N_b + KN_m$ are set to be equal during training, i.e., in one mini-batch, we used the same amount of sentence pairs (real or pseudo) and didn't need extra time for sampling. Moreover, thanks to the inherent ability of our method, when $K = 2$ we can obtain a better performance than the pseudo-NMT method in the same number of updates.

C. Low-Resource Setup

Intuitively, the amount of parallel corpus has the most significant effect against the translation quality. So, we randomly sampled 5% and 20% of the 12M bilingual sentence pairs used in En→Fr translation task and used them for pre-training respectively. We compared these smaller setups to our original 12M

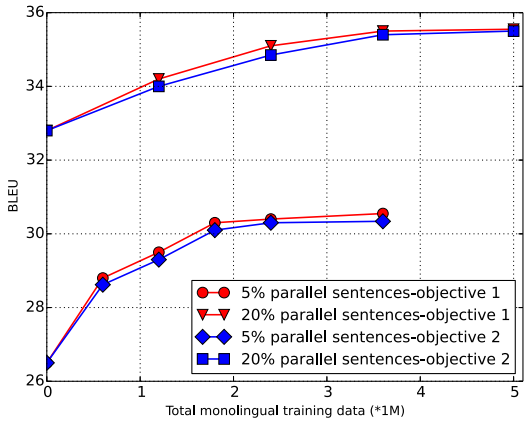


Fig. 3. BLEU scores when adding monolingual data to bitext systems with 5% and 20% of total 12M sentence pairs on En→Fr test set.

bilingual sentence pairs configuration. The results of both training objectives are shown in Figure 3. From this figure we can see that both the amount of bilingual and monolingual data have an effect on the translation quality. Specifically, when the amount of parallel data is smaller, our method makes larger improvement. This shows that the proposed mechanism makes very good utilization of monolingual data. Thus we expect that our method will be more helpful for language pairs with smaller labeled parallel data. On the other hand, for each configuration of different amount of bilingual data, we can see that the improvement tends to be flat when increasing the amount of monolingual training data, which is consistent with the results on De→En task.

D. Impact of Hyper-Parameters

There are some hyper-parameters in our proposed algorithm. In this subsection, we conducted several experiments to investigate their impact.

1) Impact of λ_1: In our proposed marginal distribution regularization objective, hyper-parameter λ_1 is introduced to balance the MLE training objective and marginal distribution regularization term in our algorithm. We conducted experiments on De→En translation to demonstrate the impact of λ_1 and plot the validation BLEU scores of different λ_1's in Figure 4(a) with respect to training iterations. From this figure, we can see that it can improve translation quality significantly and consistently against baseline with λ_1 ranging from 0.005 to 0.2, and the translation reaches the best performance when $\lambda_1 = 0.05$. Reducing or increasing λ_1 from 0.05 hurts translation quality. Similar findings are also observed on the En→Fr dataset. Therefore, we set $\lambda_1 = 0.05$ for all the experiments.

2) Impact of λ_2: For the second training objective, we introduced hyper-parameter λ_2 to balance the likelihood of bilingual data and monolingual data. To demonstrate the impact of λ_2, experiments were also conducted on De→En translation task. From Figure 4(b), we can see that the performance of the translation model reaches the peak when $\lambda_2 = 2$, and too small or too large values of λ_2 would hurt the promotion performance of adding monolingual data. We also observed similar findings on En→Fr translation, so the value of λ_2 was set as 2 for all experiments.

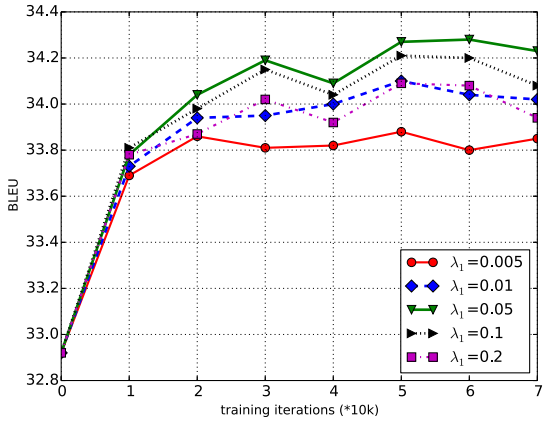
3) Impact of Sample Size K: As the inference of our approach is intractable and a plain Monte Carlo sample is highly ineffective, we propose to adopt importance sampling to sample the top-K list from distribution $P(x|y)$.

We conducted some experiments on IWSLT De→En dataset to investigate the impact of sample size K for marginal distribution regularization objective. Intuitively, a larger sample size leads to a better translation accuracy while increasing training time. To investigate the balance between translation performance and training efficiency, we trained our model with different sample sizes and present their performance. Figure 5 shows the BLEU scores of various settings of K on the validation set with respect to training hours for the marginal distribution regularization objective. From this figure, we can observe that a smaller K leads to a more rapid increase of the BLEU score on the validation set, while limiting the potential to achieve a better final accuracy. On the contrary, a larger K can achieve a better final accuracy while it takes more time to reach the good accuracy. Similar findings are also observed on the En→Fr dataset and the objective of maximizing the likelihood of both monolingual and bilingual data.

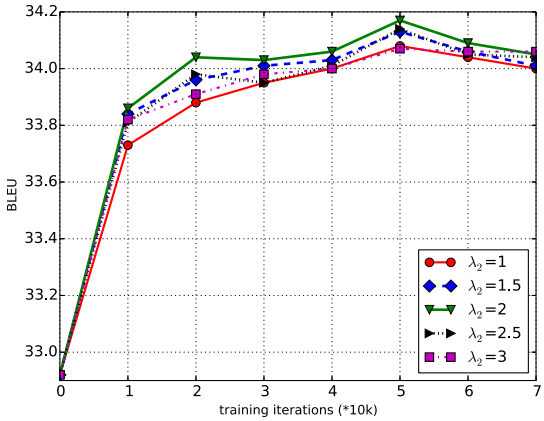
Considering the limitation of computing resource, the tradeoff between translation performance and training efficiency, as well as the fair comparison with baselines, we explored using at most five instances and eventually used two for all experiments. To be more specific, (1) a larger sample size would require larger GPU memory and longer training time, which would be unaffordable given our limited GPUs and time. (2) More samples would better approximate the marginal distribution and lead to better BLEU score, but as shown in Figure 5, the further improvements are not very significant. (3) Our approach could achieve superior results than baselines, which also use two intermediate samples as reported in the papers. Therefore, we believe that empirically, using two samples is enough to get a sufficiently good model.

E. Impact of Reverse Model for Sampling

When training model $P_\theta(y|x)$, we adopted the reverse direction model $P(x|y)$ to generate samples. We conducted several experiments with reverse direction models of different qualities on De→En translation for both training objectives. We used different En→De translation models with test BLEU score from 17.30 to 23.94 to get sampled sentences. Specifically, to generate these En→De translation models with different BLEU scores, we used the same parallel corpus as training De→En model, and selected the models with different qualities according to validation set. Finally, we selected 6 En→De translation models with BLEU scores on test set of 17.30, 19.45, 21.73, 22.06, 23.85 and 23.94, respectively. Figure 6 shows the BLEU scores of various settings of sample models on the validation set for both training objectives. From this figure, we find that using a reverse direction model $P(x|y)$ with a larger BLEU score for sampling generally leads to a better model $P_\theta(y|x)$. Therefore, we can expect a better performance when we have a reverse direction model with higher quality.



(a) Impact of λ_1 for marginal distribution regularization objective.



(b) Impact of λ_2 for the objective of maximizing the likelihood of both monolingual and bilingual data.

Fig. 4. Impact of trade-off parameters on De→En validation set.

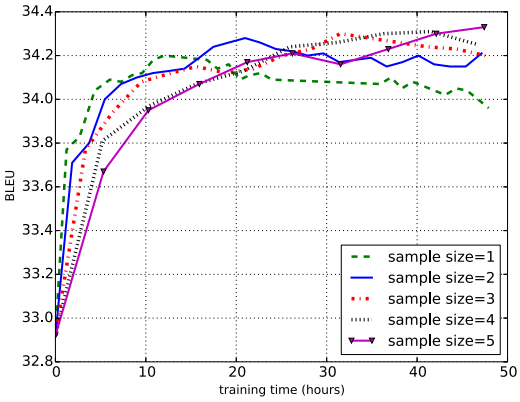


Fig. 5. Impact of sample size K on De→En validation set for marginal distribution regularization objective.

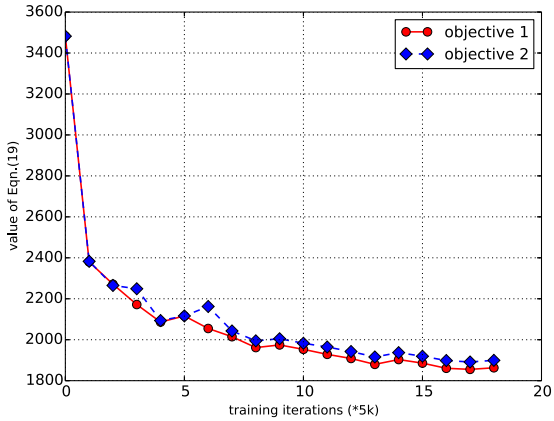


Fig. 7. The value of the regularization term on selected monolingual data during training.

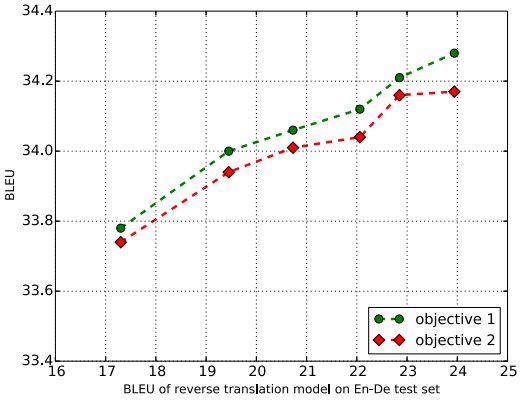


Fig. 6. Impact of reverse sample model on De→En validation set.

F. Analysis of the Regularization Term

To better understand the effect of applying the law of total probability as the regularization, we show some empirical analysis on the satisfaction of the law of total probability on monolingual data on De→En translation task for both training objectives. Specifically, after pre-training De→En translation model on parallel corpora, we randomly selected 10000 monolingual sentences and demonstrate the mean value of the following term:

$$\Big[\log \hat{P}(y') - \log \frac{1}{K}\sum_{i=1}^{K} \frac{\hat{P}(x_i)P_\theta(y'|x_i)}{P(x_i|y')}\Big]^2, \qquad (19)$$

with respect to training iterations. We plot Eqn.(19) on the selected monolingual data for both training objectives in Figure 7 with respect to training iterations. We can see that after applying both objectives to De→En translation, the value of Eqn.(19) decreases with respect to training iterations, which indicates that the marginal distribution computed by language model and estimated by importance sampling become more coherent during training. Especially, from Figure 7 we can observe that the value of Eqn.(19) for the marginal distribution regularization objective decreases more quickly compared with the other objective, which is consistent with the translation performance of the proposed objectives. For the objective of maximizing the likelihood of both bilingual and monolingual data, although we didn't enforce the law of total probability to be satisfied directly, the value of Eqn.(19) still decreases as the model becomes better, which also indicates that our assumption of the law of total

probability being satisfied for a well-trained model on any monolingual corpus is sound.

VI. Conclusion and Future Work

In this paper, we have proposed a new method to leverage monolingual corpora from a probabilistic perspective for neural machine translation. The central idea is to exploit the probabilistic connection between the marginal distribution and the conditional distribution using the law of total probability. We have introduced two different semi-supervised training objectives based on the law of total probability, including adding a data-dependent regularization term to guide the training procedure to satisfy the probabilistic connection, as well as an objective maximizing the likelihood of bilingual data and monolingual data simultaneously using the law of total probability to estimate the likelihood of monolingual data. To tackle the problem of exponentially large search space when computing the expectation term in the law of total probability, we adopted the technique of importance sampling to avoid enumerating all possible candidate source sentences and ensure the effectiveness of the proposed objectives. Experiments on English→French and German→English translation tasks show that our approach has achieved significant improvements over other semi-supervise translation approaches.

For future work, we plan to apply our method to more applications, such as speech recognition and image captioning. Furthermore, we will enrich theoretical study to better understand semi-supervised NMT via marginal distribution estimation. We will also investigate the limit of our approach with respect to the increase of the size of monolingual data as well as sample size K. Moreover, we will combine our proposed objectives with other methods, such as joint training both source-to-target and target-to-source translation models iteratively to enhance the performance of both translation models.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[2] D. Britz, A. Goldie, T. Luong, and Q. Le, "Massive exploration of neural machine translation architectures," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1442–1451.

[3] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," 2017, arXiv:1705.03122.

[4] C. Gulcehre *et al.*, "On using monolingual corpora in neural machine translation," 2015, arXiv:1503.03535.

[5] T. Mikolov, M. Karafiát, L. Burget, J. Černocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1045–1048.

[6] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 194–197.

[7] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 86–96.

[8] N. Bertoldi and M. Federico, "Domain adaptation for statistical machine translation with monolingual resources," in *Proc. 4th Workshop Statistical Mach. Transl*. 2009, pp. 182–189.

[9] P. Lambert, H. Schwenk, C. Servan, and S. Abdul-Rauf, "Investigations on translation model adaptation using monolingual data," in *Proc. 6th Workshop Statistical Mach. Transl*. 2011, pp. 284–293.

[10] N. Ueffing, G. Haffari, and A. Sarkar, "Semi-supervised model adaptation for statistical machine translation," *Mach. Transl.*, vol. 21, no. 2, pp. 77–94, 2007.

[11] D. He *et al.*, "Dual learning for machine translation," in *Proc. Advances Neural Inf. Process. Syst.*, 2016, pp. 820–828.

[12] Y. Cheng *et al.*, "Semi-supervised learning for neural machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1965–1974.

[13] Y. Wang *et al.*, "Dual transfer learning for neural machine translation with marginal distribution regularization," in *Proc. 32nd Conf. Assoc. Advancement Artif. Intell.*, 2018, pp. 5553–5560.

[14] W. He, Z. He, H. Wu, and H. Wang, "Improved neural machine translation with SMT features." in *Proc. Assoc. Advancement Artif. Intell.*, 2016, pp. 151–157.

[15] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," 2017, arXiv:1708.06426.

[16] T. Domhan and F. Hieber, "Using target-side monolingual data for neural machine translation through multi-task learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1500–1505.

[17] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," 2015, arXiv:1511.06114.

[18] P. Ramachandran, P. Liu, and Q. Le, "Unsupervised pretraining for sequence to sequence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 383–391.

[19] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 489–500.

[20] J. Zhang and C. Zong, "Exploiting source-side monolingual data in neural machine translation," in *Proc. Conf. Empirical Methods in Natural Lang. Process.*, 2016, pp. 1535–1545.

[21] L. Wu, F. Tian, T. Qin, J. Lai, and T.-Y. Liu, "A study of reinforcement learning for neural machine translation," in *Proc. Conf. Empirical Methods in Natural Lang. Process.*, 2018, pp. 3612–3621.

[22] Z. Zhang, S. Liu, M. Li, M. Zhou, and E. Chen, "Joint training for neural machine translation models with monolingual data," in *Proc. Assoc. Advancement Artif. Intell.*, 2018, pp. 555–562.

[23] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods in Natural Lang. Process.*, 2015, pp. 1412–1421.

[24] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, arXiv:1609.08144.

[25] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. V. D. Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," 2016, arXiv:1610.10099.

[26] A. Vaswani *et al.*, "Attention is all you need," 2017, arXiv:1706.03762.

[27] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," 2016, arXiv:1606.04199.

[28] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[29] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 2010.

[30] W. G. Cochran, "Sampling techniques," 3rd edn. Wiley, 1977.

[31] T. C. Hesterberg, "Advances in importance sampling," Ph.D. dissertation, Dept. Statist. Stanford University, USA, 1988.

[32] T. Hesterberg, "Weighted average importance sampling and defensive mixture distributions," *Technometrics*, vol. 37, no. 2, pp. 185–194, 1995.

[33] Y. Xia, T. Qin, W. Chen, J. Bian, N. Yu, and T.-Y. Liu, "Dual supervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3789–3798.

[34] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 11th IWSLT evaluation campaign, IWSLT 2014," in *Proc. Int. Workshop Spoken Lang. Transl.*, Hanoi, Vietnam, 2014, pp. 2–17.

[35] D. Bahdanau *et al.*, "An actor-critic algorithm for sequence prediction," 2016, arXiv:1607.07086.

[36] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. 4th Int. Conf. Learn. Representations*, 2016.

[37] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proc. 16th Conf. Eur. Assoc. Mach. Transl.*, Trento, Italy, May 2012, pp. 261–268.

[38] M. Sundermeyer, R. Schlter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. INTERSPEECH*, 2012, pp. 194–197.

[39] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980.
[40] K. Zhang *et al.*, "Drr-net: Dynamic re-read network for sentence semantic matching," in *Proc. Assoc. Advancement Artif. Intell.*, 2019.
[41] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, 2015, pp. 1–10.
[42] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, arXiv:1409.2329.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[44] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," 2015, arXiv:1508.07909.
[45] Z. Tu, Y. Liu, L. Shang, X. Liu, and H. Li, "Neural machine translation with reconstruction." in *Proc. Assoc. Advancement Artif. Intell.*, 2017, pp. 3097–3103.
[46] S. Jean, O. Firat, K. Cho, R. Memisevic, and Y. Bengio, "Montreal neural machine translation systems for WMT15," in *Proc. 10th Workshop Statistical Mach. Transl.*, 2015, pp. 134–140.
[47] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. III-1310– III-1318.
[48] M. D. Zeiler, "Adadelta: An adaptive learning rate method," 2012, arXiv:1212.5701.
[49] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. 2002, pp. 311–318.
[50] G. Lv *et al.*, "Gossiping the videos: An embedding-based generative adversarial framework for time-sync comments generation," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2019, pp. 311–318.
[51] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

Yijun Wang received the bachelor's degree in computer science from the University of Science and Technology of China, Hefei, China, in 2014. She is currently pursuing the Ph.D. degree at the School of Computer Science and Technology, University of Science and Technology of China. Her research interests include machine learning (with the focus on deep learning and reinforcement learning), artificial intelligence (with applications to neural machine translation), data mining, and recommender systems.

Yingce Xia received the bachelor's degree and Ph.D. degree both from the University of Science and Technology of China, Hefei, China. He is currently an Associate Researcher with Machine Learning Group, Microsoft Research Asia (MSRA), Beijing, China. His research interests include dual learning (a new learning paradigm proposed by our group), deep reinforcement learning (with application to neural machine translation), distributed machine learning (with application to efficient data parallel mechanism), and bandit algorithms (with special interests in budgeted MABs). He is also interested in game/auction theory.

Li Zhao received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, supervised by Professor Xiaoyan Zhu in July 2016. He is currently an Associate Researcher with Machine Learning Group, Microsoft Research Asia (MSRA), Beijing, China. Her research interests include deep learning and reinforcement learning, and their applications for text mining, finance, game, and operations research.

Jiang Bian received the bachelor's degree in computer science from Peking University, China, in 2006, and the Ph.D. degree in computer science from the Georgia Institute of Technology, Atlanta, in 2010. He is currently a Lead Researcher with Machine Learning Group, Microsoft Research Asia, Beijing, China. He has authored and coauthored more than 40 papers in refereed conferences and journals, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, International World Wide Web Conferences, ACM Knowledge Discovery and Data Mining, International Conference on Machine Learning, Annual Conference on Neural Information Processing Systems, and International Conference on Research on Development in Information Retrieval. His research interests include information retrieval, Web mining, social network analysis, and machine learning.

Tao Qin received the bachelor's degree and Ph.D. degree both from Tsinghua University, Beijing, China. He is currently a Senior Research Manager with Machine Learning Group, Microsoft Research Asia, Beijing, China. He is an Adjunct Professor (Ph.D. advisor) with the University of Science and Technology of China, Hefei, China. His research interests include machine learning (with the focus on deep learning and reinforcement learning), artificial intelligence (with applications to language understanding and computer vision), game theory and multi-agent systems (with applications to cloud computing, online and mobile advertising, ecommerce), and information retrieval and computational advertising. Prof. Qin is a member of ACM.

Enhong Chen received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China. He is a Professor and a Vice Dean with the School of Computer Science, USTC. He has authored and coauthored more than 100 papers in refereed conferences and journals, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, ACM Knowledge Discovery and Data Mining (KDD), International Conference on Data Mining (ICDM), Annual Conference on Neural Information Processing Systems, and ACM International Conference on Information and Knowledge Management. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. His research interests include data mining and machine learning, social network analysis, and recommender systems. Prof. Chen was on program committees of numerous conferences including KDD, ICDM, and SIAM International Conference on Data Mining.

Tie-Yan Liu is an Assistant Managing Director with Microsoft Research Asia, Beijing, China, leading the machine learning research area. He is an Adjunct Professor with CMU and several universities in China, and an Honorary Professor with Nottingham University. Many of his technologies have been transferred to Microsoft products and online services, such as Bing, Microsoft Advertising, Windows, Xbox, and Azure. His research interests include artificial intelligence, machine learning, information retrieval, data mining, and computational economics/finance. Prof. Liu was the recipient of many recognitions and awards in Microsoft for his significant product impacts. On the other hand, he has been actively contributing to the academic community. He is frequently invited to chair or give keynote speeches at major machine learning and information retrieval conferences. He is a distinguished member of the ACM.