The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Dual Transfer Learning for Neural Machine Translation with Marginal Distribution Regularization*

Yijun Wang,[1] Yingce Xia,[2] Li Zhao,[3] Jiang Bian,[3] Tao Qin,[3] Guiquan Liu,[1] Tie-Yan Liu[3]
[1]Anhui Province Key Lab. of Big Data Analysis and Application, University of Science and Technology of China
[2]University of Science and Technology of China
[3]Microsoft Research Asia
wyjun@mail.ustc.edu.cn, yingce.xia@gmail.com, {lizo, jiabia, taoqin, tyliu }@microsoft.com, gqliu@ustc.edu.cn

Abstract

Neural machine translation (NMT) heavily relies on parallel bilingual data 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 to learn from and regularize the training 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 dual translation model (target-to-source translation) to sample several mostly likely source-side sentences and avoid enumerating all possible candidate source sentences. That is, we transfer the knowledge contained in the dual model to boost the training of the primal model (source-to-target translation), and we call such an approach dual transfer learning. Experiment results on English→French and German→English tasks demonstrate that dual transfer learning achieves significant improvement over several strong baselines and obtains new state-of-the-art results.

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 (Bahdanau, Cho, and Bengio 2015; Britz et al. 2017; Gehring et al. 2017) 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 $P(y|x;\theta)$ 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.

(Gulcehre et al. 2015) proposed to train language models (Mikolov et al. 2010; Sundermeyer, Schlüter, and Ney 2012) 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 (Sennrich, Haddow, and Birch 2016) and statistical machine translation (Bertoldi and Federico 2009; Lambert et al. 2011; Ueffing, Haffari, and Sarkar 2007). 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.

(He et al. 2016a) propose the concept of dual learning, in which two translation models teach each other through a reinforcement learning process, by minimizing the reconstruction error of a monolingual sentences, 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, (Cheng et al. 2016) propose to append a reconstruction term to the training objective.

In this work, motivated by the law of total probability, we propose a principled way to exploit monolingual data for NMT base on transfer learning. We transfer the knowledge learned from the dual translation task (target-to-source translation) (He et al. 2016a; Xia et al. 2017b; 2017a) to our primary translation task (source-to-target translation), and we name our method as *dual transfer learning.*

*This work was conducted at Microsoft Research Asia.

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 $P(y|x;\theta)$ should satisfy the following equation:

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

However, if $P(y|x;\theta)$ is learned from bilingual corpora using maximum likelihood estimation, there is no guarantee that the above equation will hold.

Inspired by the law of total probability, we propose to learn the translation model θ by maximizing the likelihood of parallel corpora, 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.

To compute $\sum_{x\in\mathcal{X}} P(y|x;\theta)P(x)$, a technical challenge is that this value 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:

$$\begin{aligned}\sum_{x\in\mathcal{X}} P(y|x;\theta)P(x) &= \mathbb{E}_{x\sim P(x)}P(y|x;\theta) \\ &\approx \frac{1}{K}\sum_{i=1}^{K} P(y|x^{(i)};\theta), x^{(i)} \sim P(x).\end{aligned} \quad (2)$$

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 $P(y|x;\theta)$ are very sparse and most x from distribution $P(x)$ would get a nearly zero value for $P(y|x;\theta)$, 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 adopt the method of *importance sampling* and sample from distribution $P(x|y)$ to guarantee the quality of sampled sentences such that the corresponding constraint is valid empirically. Note that $P(x|y)$ is actually the dual translation model that translates a target sentence to a source sentence. Thus, by doing so, we transfer the knowledge learned from the dual translation task to our primary translation task.

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

- We propose a principled way to leverage monolingual data to enhance the training of NMT, which adopts a probabilistic view and is a kind of transfer learning.
- When estimating $\sum_{x\in\mathcal{X}} P(y|x;\theta)P(x)$, we leverage the dual translation model for importance sampling to guarantee the quality of sampled 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 baseline methods on both German→English and English→French translation tasks.

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 to a target language sentence 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 (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015; Wu et al. 2016), convolutional neural network (CNN) based frameworks (Gehring et al. 2017; Kalchbrenner et al. 2016), and, most recently, all-attention mechanisms (Vaswani et al. 2017). Beyond the standard encoder-decoder architecture, more elaborate decoder architectures have been proposed to promote the performance of NMT systems (Xia et al. 2017c; He et al. 2017). 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 (Britz et al. 2017; Zhou et al. 2016). 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.

Framework

In this section, we present a new approach, dual transfer learning, which is inspired by the law of total probability and leverages the dual translation model to learn from monolingual data. We first introduce our new training objective with a marginal distribution regularizer. Given the difficulty in estimating the regularization term brought by the exponentially large search space, we then address this challenge by using the dual model for importance sampling. After that, we present the whole dual transfer learning algorithm for NMT in details.

Training Objective

We first define some notations and present the maximum likelihood training objective used in most NMT algorithms. Then, we introduce our marginal distribution regularizer inspired by the law of total probability.

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 $\mathcal{Y}$. The translation model is usually represented by a conditional distribution $P(y|x;\theta)$ parameterized by θ, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In standard supervised learning, given a parallel corpus $\mathcal{B} = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$, the

translation model is learned by maximizing the likelihood of the training data:

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

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

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

Assume that we have a monolingual corpus $\mathcal{M}$ which contains S sentences i.i.d. sampled from the space $\mathcal{Y}$, i.e., $\mathcal{M} = \{y^{(s)}\}_{s=1}^{S}$. Considering the model $P(y|x;\theta)$ is empirically learned via maximum likelihood 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). Mathematically, we have the following constrained optimization problem:

$$\begin{aligned} &\max \sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)};\theta), \\ &s.t. P(y) = \mathbb{E}_{x\sim P(x)}P(y|x;\theta), \forall y \in \mathcal{M}. \end{aligned} \tag{5}$$

Since the ground-truth marginal distributions $P(x)$ and $P(y)$ are usually not available, we use the empirical distributions $\hat{P}(x)$ and $\hat{P}(y)$ as their surrogates, which we get from well-trained language models.

Following the common practice in constrained optimization, we convert the constraint into the following regularization term:

$$\mathcal{S}(\theta) = [\log \hat{P}(y) - \log \mathbb{E}_{x\sim\hat{P}(x)}P(y|x;\theta)]^2, \tag{6}$$

and then add it to the training objective.

Formally, we introduce our training objective as minimizing the following function:

$$\begin{aligned} \mathcal{L}(\theta) = &-\sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)};\theta) \\ &+ \lambda \sum_{s=1}^{S} [\log \hat{P}(y^{(s)}) - \log \mathbb{E}_{x\sim\hat{P}(x)}P(y^{(s)}|x;\theta)]^2, \end{aligned} \tag{7}$$

where λ is a hyperparameter controlling the tradeoff between the likelihood and the regularization term. We call this new learning scheme *maximum likelihood training with marginal distribution regularization*, since it adds a data-dependent regularization term to the original maximum likelihood training objective.

Importance Sampling with Dual Model

To compute the expectation term $\mathbb{E}_{x\sim\hat{P}(x)}P(y|x;\theta)$ in our regularizer, 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. That is, if we sample K sentences from distribution $\hat{P}(x)$, an empirical estimate of $\mathbb{E}_{x\sim\hat{P}(x)}P(y|x;\theta)$ can be computed as $\frac{1}{K}\sum_{i=1}^{K} P(y|x^i;\theta)$. However, since $P(y|x;\theta)$ is very sparse with respect to x's, most of those samples from distribution $\hat{P}(x)$ would result in $P(y|x;\theta)$ very close to zero. Intuitively, given a certain y in the target language, it is almost impossible to sample an x from empirical distribution $\hat{P}(x)$, through conforming a good source language model, such that x is exactly or close to the translation of y. In other words, most sentences sampled from $\hat{P}(x)$ are irrelevant to sentence y. Consequently, the regularization term would be constrained by nearly zero valued $P(y|x;\theta)$, which makes the constraint empirically invalid to regularize the translation model $P(y|x;\theta)$. 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.

Inspired by the ideas of dual learning (He et al. 2016a) and backtranslation (Cheng et al. 2016), we propose to get relevant source sentence x for a given target sentence y by sampling from a dual translation model $P(x|y)$. In this way, we can get constraint on $P(y|x;\theta)$ with large probability, making our constraint valid empirically. Since we sample from distribution $P(x|y)$ instead of $\hat{P}(x)$ when estimating $\mathbb{E}_{x\sim\hat{P}(x)}P(y|x;\theta)$, we need to adjust our estimate accordingly:

$$\begin{aligned} \mathbb{E}_{x\sim\hat{P}(x)}P(y|x;\theta) &= \sum_{x\in\mathcal{X}} P(y|x;\theta)\hat{P}(x) \\ &= \sum_{x\in\mathcal{X}} \frac{P(y|x;\theta)\hat{P}(x)}{P(x|y)}P(x|y) \\ &= \mathbb{E}_{x\sim P(x|y)}\frac{P(y|x;\theta)\hat{P}(x)}{P(x|y)}. \end{aligned} \tag{8}$$

That is, by making a multiplicative adjustment to $P(y|x;\theta)$, we compensate for sampling from $P(x|y)$ instead of $\hat{P}(x)$. This procedure is exactly the technique of *importance sampling* (Cochran 1977; Hesterberg 1988; 1995). Then, the importance sampling estimation of $\mathbb{E}_{x\sim\hat{P}(x)}P(y|x;\theta)$ is

$$\frac{1}{K}\sum_{i=1}^{K} \frac{P(y|x_i;\theta)\hat{P}(x_i)}{P(x_i|y)}, x_i \sim P(x|y) \tag{9}$$

where K is the sample size.

Therefore, the regularization term can be calculated approximately as follows:

$$\begin{aligned} &\mathcal{S}(\theta) \approx \\ &\sum_{s=1}^{S} \Big[\log \hat{P}(y^{(s)}) - \log \frac{1}{K}\sum_{i=1}^{K} \frac{\hat{P}(x_i^{(s)})P(y^{(s)}|x_i^{(s)};\theta)}{P(x_i^{(s)}|y^{(s)})} \Big]^2. \end{aligned} \tag{10}$$

Empirically our training objective becomes:

$$\mathcal{L}(\theta) \approx -\sum_{n=1}^{N} \log P(y^{(n)}|x^{(n)};\theta) + \lambda \sum_{s=1}^{S} \Big[\log \hat{P}(y^{(s)}) - \log \frac{1}{K} \sum_{i=1}^{K} \frac{\hat{P}(x_i^{(s)}) P(y^{(s)}|x_i^{(s)};\theta)}{P(x_i^{(s)}|y^{(s)})} \Big]^2. \quad (11)$$

Algorithm

We learn the model $P(y|x;\theta)$ by minimizing the weighted combination between the original loss function and the marginal distribution regularization term as shown in Eqn.(11). 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 from the target language B, a bilingual corpus containing sentence pairs from language A and language B, marginal distributions $\hat{P}(x)$ and $\hat{P}(y)$, and a pretrained dual model that can translate sentences from language B to language A. Denote $P(y|x;\theta)$ parameterized by θ as the translation model we want to learn and $P(x|y)$ as the dual translation model used for sampling. During training, in one mini-batch, we get m sentences form $\mathcal{M}$ and b sentence pairs from $\mathcal{B}$. Then, for each sentence y from the monolingual corpus, we sample K sentences according to the translation model $P(x|y)$. Next we compute the gradient of the objective function with respect to parameter θ and finally update the parameter θ.

Algorithm 1 Dual transfer learning with marginal distribution regularization

Require: Monolingual corpus $\mathcal{M}$, bilingual corpus $\mathcal{B}$, a dual translation model $P(x|y)$, marginal distributions $\hat{P}(x)$ and $\hat{P}(y)$, hyperparameter λ, sample size K.

1: **repeat**
2: 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 $|B_{AB}| = b$;
3: For each sentence y in M, sample K sentences $\hat{x}_1, ...\hat{x}_K$ according to the translation model $P(x|y)$;
4: Calculate the training objective $\mathcal{L}$ according to Eqn. (11) based on B_{AB}, M and the corresponding translations;
5: Update the parameters of θ:

$$\theta \leftarrow \theta - \gamma \nabla_\theta \mathcal{L}(\theta) \quad (12)$$

6: **until** model converged

Experiments

We conducted a set of experiments on two translation tasks to test the proposed method.

Settings

Datasets We evaluated our approach on two translation tasks: English→French (En→Fr) and German→English (De→En). For English→French task, we used a subset of the bilingual corpus from WMT'14 for training, which contains 12M sentence pairs. We concatenated newstest2012 and newstest2013 as the validation set, and used newstest2014 as the test set. The validation and test sets for English→French contain 6k and 3k sentence pairs respectively. We used the "News Crawl: articles from 2012" provided by WMT'14 as monolingual data. For German→English task, the bilingual corpus is from IWSLT 2014 evaluation campaign (Cettolo et al. 2014), containing about 153k sentence pairs for training, and 7k/6.5k sentence pairs for validation/test. The monolingual data for German→English is collected from web.

Baseline Methods We compared our approach with several strong baselines, including a well-known attention-based NMT system *RNNSearch* (Bahdanau, Cho, and Bengio 2015), a deep LSTM structure, and several semi-supervised NMT models:

- *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 candidate sentences obtained through beam search (Gulcehre et al. 2015).
- *Pseudo-NMT*. This method generates pseudo bilingual sentence pairs from monolingual corpora to assist training (Sennrich, Haddow, and Birch 2016). We used the same dual model to generate pseudo bilingual sentence pairs as the sampling model in our method.
- *Dual-NMT*. 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 (He et al. 2016a).

Marginal Distribution $\hat{P}(x)$ ***and*** $\hat{P}(y)$ We used LSTM-based language modeling approach to characterize the marginal distribution of a given sentence x. For En→Fr, we used a single layer LSTM with word embeddings of 512 dimensions and hidden states of 1024 dimensions. For De→En, we trained a language model with 512 dimensions for both word embeddings and hidden states. The language models were fixed during training. Both the models were trained using Adam (Kingma and Ba 2014) with initial learning rate 0.0002.

Implementation Details For En→Fr translation, we implemented a basic single-layer RNNSearch model (Bahdanau, Cho, and Bengio 2015) to ensure fair comparison with the related work, and a deep LSTM model to see improvement brought by our algorithm combining with more recent techniques. For the basic RNNSearch model, we followed the same setting as that in (Bahdanau, Cho, and Bengio 2015). To be 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 ⟨UNK⟩. For monolingual corpora, we removed the sentences containing out-of-vocabulary words. In order to prevent over-fitting, we applied dropout during training (Zaremba, Sutskever, and Vinyals 2014), where the

Table 1: BLEU scores on En→Fr and De→En translation tasks. Δ means the improvement over the basic NMT model, which only used bilingual data for training. The basic model for En→Fr is the RNNSearch model (Bahdanau, Cho, and Bengio 2015), and for De→En is a two-layer LSTM model. Note that all the methods for the same task share the same model structure.

System	En→Fr	Δ	De→En	Δ
Basic model	29.92		30.99	
Representative semi-supervised NMT systems				
Shallow fusion-NMT (Gulcehre et al. 2015)	30.03	+0.11	31.08	+0.09
Pseudo-NMT (Sennrich, Haddow, and Birch 2016)	30.40	+0.48	31.76	+0.77
Dual-NMT (He et al. 2016a)	32.06	+2.14	32.05	+1.06
Our dual transfer learning system				
This work	**32.85**	**+2.93**	**32.35**	**+1.36**

Table 2: Deep NMT systems' performances on En→Fr translation.

System	System Configurations	BLEU
Representative deep NMT systems		
(Gehring et al. 2017)	15-15 layers CNN + BPE + 12M parallel data	38.45
(Britz et al. 2017)	8-8 layers + 1024*1024 size + BPE + 36M parallel data	38.95
(Zhou et al. 2016)	9-7 layers + PosUNK +36M parallel data	39.2
Our dual transfer learning systems		
this work	4-4 layers LSTM + 512*1024 size + BPE +12M parallel data	**38.80**
	4-4 layers LSTM + 512*1024 size + BPE + 12M parallel data + Monolingual Data	**39.98**

dropout probability was 0.1. For the deep LSTM model, the dimensions of embedding and hidden states were 512 and 1024 respectively. Both the encoder and decoder had four stacked layers with residual connections (He et al. 2016b). We adopted the byte-pair encoding (BPE) techniques (Sennrich, Haddow, and Birch 2015) to split words into subwords with 32000 BPE operations, which can efficiently address rare words [1].

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

Note that our algorithm needs a dual translation model. We trained a Fr→En NMT model with test BLEU 35.46 and a De→En model with test BLEU 23.94.

Training Procedure Following (Tu et al. 2017; He et al. 2016a), to speed up training, for each task, we first trained NMT models on their own parallel corpora and then used them to initialize our algorithm.

To obtain the models used to initialize our algorithm, (1) for the single-layer RNNSearch model in English → French translation, we followed the same training procedure as that proposed by (Jean et al. 2015); (2) for deep LSTM architectures, we trained the model with mini-batch size 128 for En →Fr translation and 32 for De→En translation. Gradient clipping was used with clipping value 1.0 and 2.5 for English → French and German → English respectively. Models were optimized by AdaDelta (Zeiler 2012) on M40 GPU until convergence.

For our algorithm, we used AdaDelta with the mini-batch of 32 bilingual sentence pairs and 32 monolingual sentences for both tasks. The sample size K and the hyperparameter λ in our method were set as 2 and 0.05 respectively according to the trade-off between validation performance and training time.

Evaluation Metrics The translation qualities were measured by case-insensitive BLEU (Papineni et al. 2002) as calculated by the *multi-bleu.perl* script [2]. A larger BLEU score indicates a better translation quality. During testing, for the single-layer model in En→Fr translation, we used beam search (Sutskever, Vinyals, and Le 2014) with beam size 12 as in many previous works; for deep LSTM models, the beam size was set to 5.

Main Results

We report the experiment results in this subsection.

Table 1 shows the results of our method and three semi-supervised baselines with the aligned network structure. We can see that our dual transfer learning method outperforms all the baseline algorithms on both the language pairs. For the translation from English to French, our method outperforms the RNNSearch model with MLE training objective by 2.93 points, and outperforms the strongest baseline dual-NMT by 0.79 point. For the translation from German to English, our method outperforms the basic NMT model by 1.36 points, and outperforms dual-NMT by 0.3 points. Improvements brought by our algorithm are significant compared with the basic NMT model. These results demonstrate the effectiveness of our algorithm.

Table 2 shows the comparison between our proposed algorithm and several deep NMT systems on the En→Fr translation task. We can see that given a strong baseline, our al-

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

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

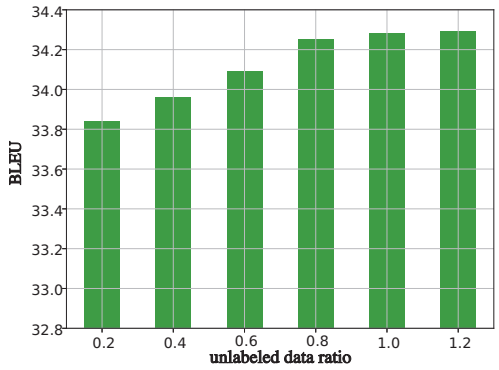

Figure 1: Impact of unlabeled data ratio on German→ English validation set.

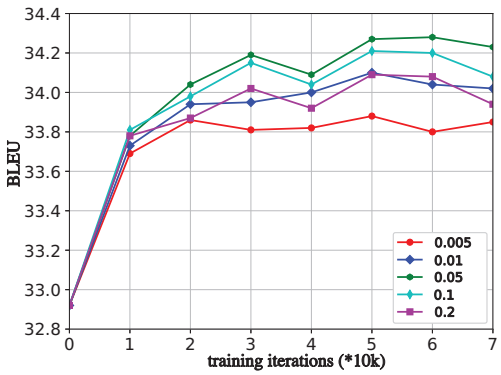

Figure 2: Impact of λ on German→English validation set.

gorithm can still make significant improvement, i.e., from 38.80 to 39.98. This sets a new record on En→Fr translation with 12M bilingual data. We leave leveraging more bilingual/monolingual data as a future work.

Given a parallel corpus, one may be curious about that how many unlabeled sentences are most beneficial to improve 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 divided by the number of labeled sentence pairs in each mini-batch. Figure 1 shows the BLEU scores of the German→English 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 improvement on translation quality, while the improvement tends to be marginal if further increasing the ratio. Therefore, considering the balance between model performance and training time, we set the ratio to 1 in all other experiments.

Impact of hyperparameters

There are some hyperparameters in our marginal distribution regularization algorithm. In this subsection, we conducted several experiments to investigate their impact.

Impact of λ Hyperparameter λ is introduced to balance the MLE training objective and the regularization term in our algorithm. We conducted experiments on German→English translation to study the impact of λ. We plot the validation BLEU scores of different λ's in Figure 2 with respect to training iterations. From this figure, we can see that $\lambda \in [0.005, 0.2]$ can improve translation quality significantly and consistently against baseline, and $\lambda = 0.05$ reaches the best performance. Reducing or increasing λ from 0.05 hurts translation quality. Similar findings are also observed on the English→French dataset. Therefore, we set $\lambda = 0.05$ for all the experiments.

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 use the dual model to sample the top-K list from distribution $P(x|y;\theta_{y\to x})$.

We conducted some experiments on IWSLT German→English dataset to study the impact of sample size K. 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. Figure 3 shows the BLEU scores of various settings of K on the validation set with respect to training hours. 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 higher final accuracy. On the contrary, a larger K achieves a higher final accuracy while taking more time to reach the good accuracy. Similar findings are also observed on the En→Fr dataset. Due to limited computation resources, we set $K = 2$ in all experiments.

Impact of the dual model for sampling When training model $P(y|x;\theta)$, we adopted the dual translation model $P(x|y)$ to generate samples. We conducted several experiments with dual models of different qualities on German→English translation. We used different En→De translation models with test BLEU score from 17.30 to 23.94 to sample sentences. As can be seen from Figure 4, using a dual model $P(x|y)$ with a larger BLEU score for sampling generally leads to higher final accuracy. Therefore, we expect we can further improve the accuracy if we are give a better dual model.

Related Work

Exploring monolingual data for machine translation has attracted intensive attention in recent years. 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 separately trained language model with monolingual data is integrated into the NMT model.

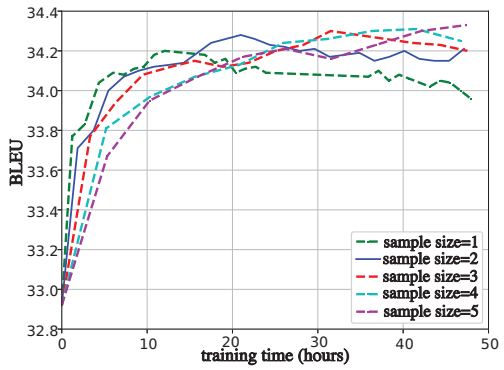

Figure 3: Impact of sample size K on German→English validation set.

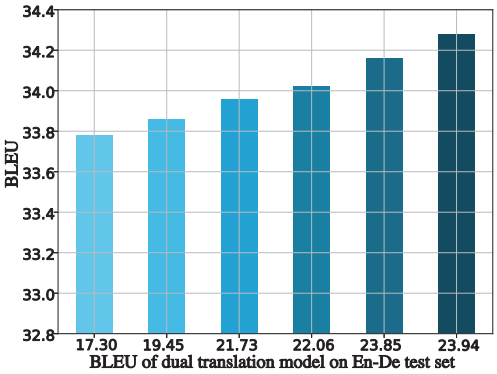

Figure 4: Impact of the dual translation model on German→English validation set.

(Gulcehre et al. 2015) trained language models independently with target-side monolingual sentences, and incorporated them 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. (Jean et al. 2015) also reported experiments of reranking NMT outputs with a 5-gram language model. These methods only used monolingual data to train language models and improve NMT decoding, but do not touch the training of NMT models.

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, (Bertoldi and Federico 2009; Lambert et al. 2011) have back-translated target-side monolingual data into the source-side sentence to produce synthetic parallel data for phrase-based SMT. 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 (Sennrich, Haddow, and Birch 2016). (Ueffing, Haffari, and Sarkar 2007) iteratively translated source-side monolingual data and added the reliable translations to the training data in an SMT system, and thus improved the translation model from its own translation. For these methods, there is no guarantee on the quality of generated pseudo bilingual sentence pairs, which may limit the performance gain.

In the third category, the monolingual data is reconstructed with both source-to-target and target-to-source translation models, and the two models are jointly trained. (He et al. 2016a) proposed dual learning for NMT, in which two translation models taught each other through a reinforcement learning process, based on the feedback signals generated during this process. (Cheng et al. 2016) proposed to append a reconstruction term to the training objective, which aims to reconstruct the observed monolingual corpora using an autoencoder. To some extent, the reconstruction methods could be seen as an iteration extension of (Sennrich, Haddow, and Birch 2016)'s method, since after updating model parameters on the pseudo parallel corpus, the learned models are used to produce a better pseudo corpus (Cheng et al. 2016). Different form those methods, which focus on reconstruction of monolingual sentences, our approach focuses on the endogenous probabilistic connection between the marginal distribution of monolingual data and the conditional distribution represented by the translation model. To some extent, our approach is a more principled way.

Transfer learning is a broad research direction in machine learning. Different from most transfer learning methods (Raina et al. 2007; Long et al. 2015; 2016), our algorithm leverages the dual structure of machine translation and achieves knowledge transfer through data sampling.

Conclusion

In this paper, we have proposed a new method, dual transfer learning, 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. A data-dependent regularization term is introduced to guide the training procedure to satisfy the probabilistic connection. The key technical challenge is addressed by using the dual translation model for important sampling. Experiments on English→French and German→English translation tasks show that our approach has achieved significant improvements over baseline methods.

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 dual transfer learning with marginal distribution regularization. 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.

Acknowledgements

This research was partially supported by grants from the National Key Research and Development Program of China (Grant No.2016YFB1000904), and the National Natural Science Foundation of China (Grants No.61727809).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. *ICLR*.

Bertoldi, N., and Federico, M. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the fourth workshop on statistical machine translation*, 182–189. Association for Computational Linguistics.

Britz, D.; Goldie, A.; Luong, T.; and Le, Q. 2017. Massive exploration of neural machine translation architectures. *ACL*.

Cettolo, M.; Niehues, J.; Stüker, S.; Bentivogli, L.; and Federico, M. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*.

Cheng, Y.; Xu, W.; He, Z.; He, W.; Wu, H.; Sun, M.; and Liu, Y. 2016. Semi-supervised learning for neural machine translation. *arXiv preprint arXiv:1606.04596*.

Cochran, W. G. 1977. Sampling techniques. *John Wiley*.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Gulcehre, C.; Firat, O.; Xu, K.; Cho, K.; Barrault, L.; Lin, H.-C.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.; and Ma, W.-Y. 2016a. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, 820–828.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, D.; Lu, H.; Xia, Y.; Qin, T.; Wang, L.; and Liu, T.-Y. 2017. Decoding with value networks for neural machine translation. In *Advances in Neural Information Processing Systems*.

Hesterberg, T. C. 1988. *Advances in importance sampling*. Ph.D. Dissertation, Stanford University.

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

Jean, S.; Firat, O.; Cho, K.; Memisevic, R.; and Bengio, Y. 2015. Montreal neural machine translation systems for wmt15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 134–140.

Kalchbrenner, N.; Espeholt, L.; Simonyan, K.; Oord, A. v. d.; Graves, A.; and Kavukcuoglu, K. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lambert, P.; Schwenk, H.; Servan, C.; and Abdul-Rauf, S. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 284–293. Association for Computational Linguistics.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 97–105.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, 136–144.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *EMNLP*.

Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, 3.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. Y. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, 759–766. ACM.

Sennrich, R.; Haddow, B.; and Birch, A. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving neural machine translation models with monolingual data. *Annual Meeting of the Association for Computational Linguistics* 11–11.

Sundermeyer, M.; Schlüter, R.; and Ney, H. 2012. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Tu, Z.; Liu, Y.; Shang, L.; Liu, X.; and Li, H. 2017. Neural machine translation with reconstruction. In *AAAI*, 3097–3103.

Ueffing, N.; Haffari, G.; and Sarkar, A. 2007. Semi-supervised model adaptation for statistical machine translation. *Machine Translation* 21(2):77–94.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; N. Gomez, A.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xia, Y.; Bian, J.; Qin, T.; Yu, N.; and Liu, T.-Y. 2017a. Dual inference for machine learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3112–3118.

Xia, Y.; Qin, T.; Chen, W.; Bian, J.; Yu, N.; and Liu, T.-Y. 2017b. Dual supervised learning. In *International Conference on Machine Learning*, 3789–3798.

Xia, Y.; Tian, F.; Wu, L.; Lin, J.; Qin, T.; and Liu, T.-Y. 2017c. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*.

Zaremba, W.; Sutskever, I.; and Vinyals, O. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

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