

Dual Reconstruction: a Unifying Objective for Semi-Supervised Neural Machine Translation

Weijia Xu

University of Maryland
weijia@cs.umd.edu

Xing Niu*

Amazon AI
xingniu@amazon.com

Marine Carpuat

University of Maryland
marine@cs.umd.edu

Abstract

While Iterative Back-Translation and Dual Learning effectively incorporate monolingual training data in neural machine translation, they use different objectives and heuristic gradient approximation strategies, and have not been extensively compared. We introduce a novel dual reconstruction objective that provides a unified view of Iterative Back-Translation and Dual Learning. It motivates a theoretical analysis and controlled empirical study on German-English and Turkish-English tasks, which both suggest that Iterative Back-Translation is more effective than Dual Learning despite its relative simplicity.

1 Introduction

Taking advantage of monolingual training data via Back-Translation (Sennrich et al., 2016a), Iterative Back-Translation (Zhang et al., 2018; Cotterell and Kreutzer, 2018) or Dual Learning (He et al., 2016) has become a de facto requirement for building high quality Neural Machine Translation (NMT) systems (Edunov et al., 2018; Hassan et al., 2018). However, these methods rely on unrelated heuristic optimization objectives, and it is not clear what their respective strengths and weaknesses are, nor how they relate to the ideal but intractable objective of maximizing the marginal likelihood of the monolingual data (i.e., $p_\theta(\mathbf{y}) = \sum_{\mathbf{x}} p_\theta(\mathbf{y} | \mathbf{x}) q(\mathbf{x})$) given target sentences \mathbf{y} , an NMT model $p_\theta(\mathbf{y} | \mathbf{x})$, and the prior distribution $q(\mathbf{x})$ on source \mathbf{x}).

Instead of proposing new methods, this paper sheds new light on how these established techniques work and how to use them. We introduce a **dual reconstruction objective** to theoretically ground the comparison of semi-supervised training strategies that leverage monolingual data from both source and target languages (Figure 1). In Section 3, we show that, under some assumptions, this

*Work was done at the University of Maryland.

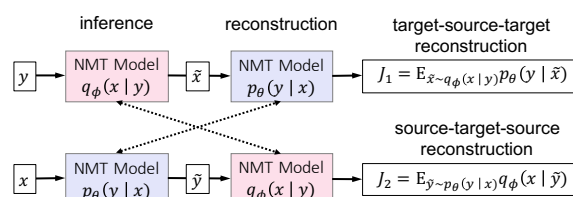


Figure 1: Our dual reconstruction objective sums 1) a target-source-target objective \mathcal{J}_1 on target sentences \mathbf{y} using the NMT model $q_\phi(\mathbf{x} | \mathbf{y})$ for inference and $p_\theta(\mathbf{y} | \mathbf{x})$ for reconstruction, and 2) a source-target-source objective \mathcal{J}_2 on source sentences \mathbf{x} using $p_\theta(\mathbf{y} | \mathbf{x})$ for inference and $q_\phi(\mathbf{x} | \mathbf{y})$ for reconstruction. Models connected by dotted arrows share parameters.

objective remarkably shares the same global optimum as the intractable marginal likelihood objective where the model’s marginal distribution $p_\theta(\mathbf{y})$ coincides with the target sentence distribution $p(\mathbf{y})$. We also show that Iterative Back-Translation (IBT) and Dual Learning can be viewed as different ways to approximate its optimization.

Theory suggests that IBT approximates the dual reconstruction objective more closely than the more complex Dual Learning approach, and in particular that Dual Learning’s additional language model loss is redundant. We investigate whether these differences matter in practice by conducting the first controlled empirical comparison of Back-Translation, IBT, and Dual Learning in high-resource (WMT de-en), low-resource (WMT tr-en), and cross-domain settings (News→TED, de-en). Results support our theory that the additional language model loss and policy gradient estimation in Dual Learning is redundant and show that IBT outperforms the more complex Dual Learning algorithm in terms of translation quality. Furthermore, we also compare different optimization strategies used in IBT to better balance translation quality against the computational cost.

2 Background

Notation NMT models the probability of translating a source sequence \mathbf{x} into a target \mathbf{y} as $p_\theta(\mathbf{y} | \mathbf{x}) = \prod_{t=1}^T p(y_t | \mathbf{y}_{<t}, \mathbf{x}; \theta)$ where θ represents the model parameters, and T is the length of \mathbf{y} (Bahdanau et al., 2015). The model computes the conditional probability of the next token at time t by $p(\cdot | \mathbf{y}_{<t}, \mathbf{x}; \theta) = \text{softmax}(a(\mathbf{h}_t))$, where $a(\cdot)$ is a linear transformation, and \mathbf{h}_t is the hidden representation at step t usually modeled by an encoder-decoder network $\mathbf{h}_t = f(\mathbf{y}_{<t}, \mathbf{x})$. In supervised settings, NMT models are trained to maximize the likelihood of parallel sentence pairs: $\mathcal{J}_s(\theta) = \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} \log p_\theta(\mathbf{y} | \mathbf{x})$ given data $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^N$.

IBT and Dual Learning exploit large monolingual corpora which represent source and target language distributions better than the limited parallel corpora. Back-Translation trains the source-to-target translation model $p_\theta(\mathbf{y} | \mathbf{x})$ by maximizing the conditional log-likelihood of target language sentences \mathbf{y} given pseudo source sentences $\tilde{\mathbf{x}}$ inferred by a pre-trained target-to-source translation model $q_\phi(\mathbf{x} | \mathbf{y})$ given \mathbf{y} . IBT optimizes the dual translation models $p_\theta(\mathbf{y} | \mathbf{x})$ and $q_\phi(\mathbf{x} | \mathbf{y})$ via back-translation in turn, both for semi-supervised (Zhang et al., 2018; Hoang et al., 2018; Cotterell and Kreutzer, 2018; Niu et al., 2018) and unsupervised MT (Artetxe et al., 2018; Lample et al., 2018a,b). Dual Learning takes the view of cooperative game theory where dual models collaborate with each other to learn to reconstruct the observed source and target monolingual sentences, and is widely used for semi-supervised (He et al., 2016), unsupervised (Wang et al., 2019), and zero-shot multilingual NMT (Sestorain et al., 2018). Concretely, Dual Learning optimizes $p_\theta(\mathbf{y} | \mathbf{x})$ and $q_\phi(\mathbf{x} | \mathbf{y})$ jointly by reconstructing the original target sentence \mathbf{y} using $p_\theta(\mathbf{y} | \mathbf{x})$ given the source $\tilde{\mathbf{x}}$ inferred by $q_\phi(\mathbf{x} | \mathbf{y})$, and vice versa. The reconstruction loss is augmented with a language model loss and used to update both reconstruction and inference models via policy gradient (Williams, 1992).

While Dual Learning and IBT each improve BLEU over Back-Translation (Zhang et al., 2018; Cotterell and Kreutzer, 2018; He et al., 2016), they have not been compared directly to each other. Cotterell and Kreutzer (2018) interpret Back-Translation as a variational approximation where the pseudo source $\tilde{\mathbf{x}}$ can be viewed as a latent vari-

able and the target-to-source model $q_\phi(\mathbf{x} | \mathbf{y})$ is an inference network that approximates the posterior distribution $p_\theta(\mathbf{x} | \mathbf{y})$. Furthermore, they explain IBT as a way to better approximate the true posterior distribution with the target-to-source model. However, it is unclear how their heuristic objective relates to the ideal objective of maximizing the model’s marginal likelihood of the target language monolingual data. More recently, He et al. (2020) connect back-translation and the language model loss in Dual Learning to the variational lower-bound (ELBO) of the marginal likelihood objective. We introduce a more direct connection.

3 Theoretical View with Dual Reconstruction Objective

3.1 Variational Auto-Encoders for Semi-Supervised MT

Following Cotterell and Kreutzer (2018), we define a generative latent variable model of bitext

$$p_\theta(\mathbf{x}, \mathbf{y}) = p_\theta(\mathbf{y} | \mathbf{x})q(\mathbf{x})$$

where the source \mathbf{x} is randomly sampled from the prior distribution $q(\mathbf{x})$ estimated by the empirical data distribution $q_{data}(\mathbf{x})$ based on the abundant source monolingual data $\mathcal{M}_X = \{\mathbf{x}^{(m)}\}_{m=1}^M$:

$$q_{data}(\mathbf{x}) = \begin{cases} \frac{1}{|\mathcal{M}_X|}, & \text{if } \mathbf{x} \in \mathcal{M}_X \\ 0, & \text{otherwise} \end{cases}$$

and the target translation \mathbf{y} is sampled from the translation model $p_\theta(\mathbf{y} | \mathbf{x})$ conditioned on \mathbf{x} .

Given the target sentence distribution $p(\mathbf{y})$ estimated by the empirical data distribution $p_{data}(\mathbf{y})$ of target monolingual data $\mathcal{M}_Y = \{\mathbf{y}^{(m)}\}_{m=1}^{M'}$, we can view \mathbf{x} as a latent variable and maximize the marginal log-likelihood

$$\mathcal{J}_u(\theta) = \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} [\log p_\theta(\mathbf{y})]$$

where $p_\theta(\mathbf{y})$ is the model’s marginal likelihood $p_\theta(\mathbf{y}) = \sum_{\mathbf{x}} p_\theta(\mathbf{x}, \mathbf{y})$. The global optimum of the objective is achieved when the model’s marginal distribution $p_\theta(\mathbf{y})$ perfectly matches the target sentence distribution $p(\mathbf{y})$.¹

However, directly optimizing the marginal likelihood $p_\theta(\mathbf{y})$ is intractable due to the infinite space

¹We will define constraints to guarantee avoiding the uninteresting solution where $p_\theta(\mathbf{y} | \mathbf{x}) = p(\mathbf{y})$ in Section 3.2.

of \mathbf{x} . We can instead apply variational auto-encoding (VAE) models by introducing an inference network $p_\psi(\mathbf{x} | \mathbf{y})$ and maximize the variational lower-bound (ELBO) of $\log p_\theta(\mathbf{y})$:

$$\log p_\theta(\mathbf{y}) \geq \mathbb{E}_{\mathbf{x} \sim p_\psi(\mathbf{x} | \mathbf{y})} [\log p_\theta(\mathbf{y} | \mathbf{x})] - D_{\text{KL}} [p_\psi(\mathbf{x} | \mathbf{y}) || q(\mathbf{x})] \quad (1)$$

where $D_{\text{KL}} [p_\psi || q]$ is the Kullback-Leibler (KL) divergence. However, estimating the prior distribution $q(\mathbf{x})$ by the discrete data distribution $q_{\text{data}}(\mathbf{x})$ makes it difficult to directly compute the KL term. One can estimate $q(\mathbf{x})$ using a language model (LM) trained to maximize the likelihood of the source monolingual data (Miao and Blunsom, 2016; Baziotis et al., 2019), at the cost of introducing additional model bias into the translation model. The non-differentiable KL term requires gradient estimators such as policy gradient (Williams, 1992) or Gumbel-softmax (Jang et al., 2017), which may introduce further training noise (He et al., 2020).

To address these issues, we introduce the dual reconstruction objective, which includes two reconstruction terms that resemble the first term in the ELBO objective (Eq. (1)) while excluding the KL term that is challenging to optimize and show that this objective has desirable properties and can be better approximated in practice.

Definition 3.1. Given prior distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ over the sentences \mathbf{x} in the source language space Σ_x and \mathbf{y} in the target language space Σ_y , we define the **dual reconstruction objective** $\mathcal{J}_{\text{dual}}(\theta, \phi)$ for dual translation models $p_\theta(\mathbf{y} | \mathbf{x})$ and $q_\phi(\mathbf{x} | \mathbf{y})$ as the sum of the target-source-target objective \mathcal{J}_1 and source-target-source objective \mathcal{J}_2 :

$$\begin{aligned} \mathcal{J}_{\text{dual}}(\theta, \phi) &= \mathcal{J}_1(\theta, \phi) + \mathcal{J}_2(\theta, \phi) \\ \mathcal{J}_1(\theta, \phi) &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} [\log p_\theta(\mathbf{y} | \mathbf{x})] \right] \\ \mathcal{J}_2(\theta, \phi) &= \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \left[\mathbb{E}_{\mathbf{y} \sim p_\theta(\mathbf{y} | \mathbf{x})} [\log q_\phi(\mathbf{x} | \mathbf{y})] \right] \end{aligned} \quad (2)$$

For \mathcal{J}_1 , the target-to-source model $q_\phi(\mathbf{x} | \mathbf{y})$ serves as the **inference model** to produce pseudo source sequences $\tilde{\mathbf{x}}$ given target sequences \mathbf{y} and $p_\theta(\mathbf{y} | \mathbf{x})$ serves as the **reconstruction model** to reconstruct \mathbf{y} given $\tilde{\mathbf{x}}$, and vice versa for \mathcal{J}_2 . We first define the mutual information constraint in Section 3.2 and show in Section 3.3 that $\mathcal{J}_{\text{dual}}(\theta, \phi)$ shares the same global optimum as the marginal

likelihood objective which is intractable to optimize directly.² In Section 3.4, we compare and contrast how IBT and Dual Learning approximate $\mathcal{J}_{\text{dual}}(\theta, \phi)$.

3.2 Mutual Information Constraint

The global optimum of the marginal likelihood objective is achieved when the model’s marginal distribution $p_\theta(\mathbf{y}) = p(\mathbf{y})$. Given a translation model with enough capacity without any constraint on how the model output is dependent on the source context, this could lead to a degenerate solution $p_\theta(\mathbf{y} | \mathbf{x}) = p(\mathbf{y})$ where the model ignores the source input and memorizes the monolingual training data. We constrain the translation model to avoid this situation, using the mutual information of a conditional distribution $p_\theta(\mathbf{y} | \mathbf{x})$ which measures how much \mathbf{y} is dependent on \mathbf{x} in p_θ (Hoffman and Johnson, 2016). Here, this mutual information measures the degree to which model translations depend on the source.

Definition 3.2. Given a prior distribution $q(\mathbf{x})$ over $\mathbf{x} \in \Sigma_x$, we define the **mutual information** I_{p_θ} of \mathbf{x} and \mathbf{y} in the conditional distribution $p_\theta(\mathbf{y} | \mathbf{x})$:

$$I_{p_\theta} = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} [D_{\text{KL}} [p_\theta(\mathbf{y} | \mathbf{x}) || p_\theta(\mathbf{y})]] \quad (3)$$

where $p_\theta(\mathbf{y})$ is the marginal distribution:

$$p_\theta(\mathbf{y}) = \sum_{\mathbf{x}} p_\theta(\mathbf{y} | \mathbf{x}) q(\mathbf{x}) \quad (4)$$

To avoid the degenerate solution, we constrain the model’s mutual information by:

$$0 \leq I_{\text{min}} \leq I_{p_\theta} \leq I_{\text{max}} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y})$$

where I_{min} and I_{max} are pre-defined constant values between zero and the maximum mutual information between \mathbf{x} and \mathbf{y} given any joint distribution $p(\mathbf{x}, \mathbf{y}) \in P_{XY}$ whose marginals satisfy $\sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})$ and $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})$. Hledík et al. (2019) prove that the maximum mutual information $\max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y}) = \min(\mathbb{H}[q(\mathbf{x})], \mathbb{H}[p(\mathbf{y})])$, where $\mathbb{H}[q(\mathbf{x})]$ and $\mathbb{H}[p(\mathbf{y})]$ are the entropy of prior distributions $q(\mathbf{x})$ and $p(\mathbf{y})$. Thus, the maximum mutual information should be large enough to properly bound the model’s mutual information

²We focus on key components of the proof and leave detailed derivations for supplemental material.

if $q(\mathbf{x})$ and $p(\mathbf{y})$ are defined on large monolingual corpora \mathcal{M}_X and \mathcal{M}_Y .

Intuitively, the constraint requires that the model's mutual information cannot be so small that the model ignores the source context nor so large such that is not robust to the noise in the source input. We will show in Section 4.4 that in practice, this constraint is met when jointly optimizing the supervised and unsupervised objectives without explicitly applying constrained optimization.

3.3 Understanding the Global Optimum of the Dual Reconstruction Objective

We first characterize the upper bound of the dual reconstruction objective.

Proposition 1. Given prior distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ over $\mathbf{x} \in \Sigma_x$ and $\mathbf{y} \in \Sigma_y$, if parameterized probability models p_θ and q_ϕ have enough capacity under the constraint that:

$$0 \leq I_{min} \leq I_{p_\theta}, I_{q_\phi} \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y})$$

where I_{min} and I_{max} are pre-defined constant values between zero and the maximum mutual information between \mathbf{x} and \mathbf{y} given any joint distribution $p(\mathbf{x}, \mathbf{y}) \in P_{XY}$ whose marginals satisfy $\sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})$ and $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})$. Then, the dual reconstruction objective is upper-bounded by $\mathcal{J}_{dual}(\theta, \phi) \leq 2I_{max} - \mathbb{H}[q(\mathbf{x})] - \mathbb{H}[p(\mathbf{y})]$, and the upper bound is achieved iff

$$\begin{aligned} I_{q_\phi} &= I_{max} \\ I_{p_\theta} &= I_{max} \\ p_\theta(\mathbf{y} | \mathbf{x}) &= \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y}) \\ q_\phi(\mathbf{x} | \mathbf{y}) &= \frac{p_\theta(\mathbf{y} | \mathbf{x})}{p_\theta(\mathbf{y})} q(\mathbf{x}) \end{aligned} \quad (5)$$

Proof. First we prove that $\mathcal{J}_1(\theta, \phi) \leq I_{max} - \mathbb{H}[p(\mathbf{y})]$, and the upper bound is achieved iff

$$\begin{aligned} I_{q_\phi} &= I_{max} \\ p_\theta(\mathbf{y} | \mathbf{x}) &= \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y}) \end{aligned}$$

To show this, we denote the posterior distribution $Q(\mathbf{y} | \mathbf{x}) = \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y})$, and rewrite \mathcal{J}_1 :

$$\begin{aligned} \mathcal{J}_1 &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} [\log p_\theta(\mathbf{y} | \mathbf{x})] \right] \\ &= I_{q_\phi} - \mathbb{H}[p(\mathbf{y})] \\ &\quad - D_{KL} [q_\phi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) || p_\theta(\mathbf{y} | \mathbf{x}) q_\phi(\mathbf{x})] \end{aligned}$$

Since the KL divergence between two distributions is always non-negative and is zero iff they are equal, we have

$$\mathcal{J}_1(\theta, \phi) \leq I_{q_\phi} - \mathbb{H}[p(\mathbf{y})] \leq I_{max} - \mathbb{H}[p(\mathbf{y})]$$

and $\mathcal{J}_1(\theta, \phi) = I_{max} - \mathbb{H}[p(\mathbf{y})]$ iff

$$\begin{aligned} I_{q_\phi} &= I_{max} \\ D_{KL} [q_\phi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) || p_\theta(\mathbf{y} | \mathbf{x}) q_\phi(\mathbf{x})] &= 0 \end{aligned}$$

The second equality holds iff

$$p_\theta(\mathbf{y} | \mathbf{x}) = \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y})$$

Similarly, we can prove that $\mathcal{J}_2(\theta, \phi) \leq I_{max} - \mathbb{H}[q(\mathbf{x})]$, and the upper bound is achieved iff

$$\begin{aligned} I_{p_\theta} &= I_{max} \\ q_\phi(\mathbf{x} | \mathbf{y}) &= \frac{p_\theta(\mathbf{y} | \mathbf{x})}{p_\theta(\mathbf{y})} q(\mathbf{x}) \end{aligned}$$

thus $\mathcal{J}_{dual}(\theta, \phi) \leq 2I_{max} - \mathbb{H}[q(\mathbf{x})] - \mathbb{H}[p(\mathbf{y})]$ and the upper bound is achieved iff θ and ϕ satisfy Eq. (5), concluding the proof. \square

Proposition 1 shows that $\mathcal{J}_{dual}(\theta, \phi)$ has an upper bound that could be reached when the mutual information of $p_\theta(\mathbf{y} | \mathbf{x})$ and $q_\phi(\mathbf{x} | \mathbf{y})$ are maximized, and $p_\theta(\mathbf{y} | \mathbf{x})$ and $q_\phi(\mathbf{x} | \mathbf{y})$ are equal to the posterior distribution for each other. Next we show that the upper bound is indeed the global maximum of the objective $\mathcal{J}_{dual}(\theta, \phi)$, as there exists a solution for the above conditions (proof in Appendix A.2).

Proposition 2. Given distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ over $\mathbf{x} \in \Sigma_x$ and $\mathbf{y} \in \Sigma_y$, if parameterized probability models p_θ and q_ϕ have enough capacity under the constraint that:

$$0 \leq I_{min} \leq I_{p_\theta}, I_{q_\phi} \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y}) \quad (6)$$

where I_{min} and I_{max} are pre-defined constant values between zero and the maximum mutual information between \mathbf{x} and \mathbf{y} given any joint distribution $p(\mathbf{x}, \mathbf{y}) \in P_{XY}$ whose marginals satisfy $\sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})$ and $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})$. Then there exist θ^* and ϕ^* such that:

$$\begin{aligned} I_{q_{\phi^*}} &= I_{p_{\theta^*}} = I_{max} \\ p_{\theta^*}(\mathbf{y} | \mathbf{x}) &= \frac{q_{\phi^*}(\mathbf{x} | \mathbf{y})}{q_{\phi^*}(\mathbf{x})} p(\mathbf{y}) \\ q_{\phi^*}(\mathbf{x} | \mathbf{y}) &= \frac{p_{\theta^*}(\mathbf{y} | \mathbf{x})}{p_{\theta^*}(\mathbf{y})} q(\mathbf{x}) \end{aligned} \quad (7)$$

Finally, we connect the global optimum of the dual reconstruction objective to that of the marginal likelihood objective (proof in Appendix A.3).

Theorem 1. Given prior distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ over $\mathbf{x} \in \Sigma_x$ and $\mathbf{y} \in \Sigma_y$, if parameterized probability models p_θ and q_ϕ have enough capacity under the constraint that:

$$0 \leq I_{min} \leq I_{p_\theta}, I_{q_\phi} \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y})$$

where I_{min} and I_{max} are pre-defined constant values between zero and the maximum mutual information between \mathbf{x} and \mathbf{y} given any joint distribution $p(\mathbf{x}, \mathbf{y}) \in P_{XY}$ whose marginals satisfy $\sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})$ and $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})$. Let θ^*, ϕ^* be the global optimum of the dual reconstruction objective $\max_{\theta, \phi} \mathcal{J}_{dual}(\theta, \phi)$, then $q_{\phi^*}(\mathbf{x}) = q(\mathbf{x})$, $p_{\theta^*}(\mathbf{y}) = p(\mathbf{y})$, and $I_{q_{\phi^*}} = I_{p_{\theta^*}} = I_{max}$.

Thus, while the marginal likelihood objective provides no guarantee for the model’s mutual information, the global optimum of dual reconstruction objective guarantees that the mutual information of translation models $p_\theta(\mathbf{y} | \mathbf{x})$ and $q_\phi(\mathbf{x} | \mathbf{y})$ will be maximized to I_{max} .

3.4 Practical Approximations

Despite its desirable optimum, the dual reconstruction objective cannot be directly optimized since decoding is not differentiable. We compare how it is approximated by IBT vs. Dual Learning.

Gradient Approximation To estimate the dual reconstruction objective, one could use sampling or beam search from the model distribution. However, since neither approach is differentiable, the gradients $\nabla_{\theta} \mathcal{J}_2$ and $\nabla_{\phi} \mathcal{J}_1$ cannot be computed directly. IBT blocks the gradients $\nabla_{\theta} \mathcal{J}_2$ and $\nabla_{\phi} \mathcal{J}_1$ assuming that they are negligible, while Dual Learning approximates them by policy gradient (Williams, 1992), which can lead to slow and unstable training (Henderson et al., 2018; Wu et al., 2018). Proposition 1 shows that the objective is maximized when the mutual information is maximized to I_{max} . Thus, maximizing the mutual information by other means can help side-step this issue. For example, combining the supervised and unsupervised training objectives (Sennrich et al., 2016a; Cotterell and Kreutzer, 2018) to train models jointly on the parallel and monolingual data can help. For unsupervised MT, the denoising auto-encoding objective introduced in Lample et al. (2018a) can be viewed as a way to maximize the mutual information.

LM Loss Dual Learning combines the dual reconstruction objective with an LM loss to encourage the generated translations to be close to the target language domain. Theorem 1 suggests that the LM loss is redundant: optimizing the dual reconstruction objective implicitly pushes the output distributions of the source-to-target and target-to-source models toward the target and source language distributions respectively, which has the same effect intended by the LM loss.

Optimization Strategy While Dual Learning uses batch-level updates, where back-translations are generated on-the-fly and the translation models p_θ and q_ϕ are updated alternately in data batches, IBT adopts different strategies based on the data settings. Batch-level IBT is used in unsupervised MT to quickly boost the model performance from a cold start (Artetxe et al., 2018; Lample et al., 2018a), while epoch-level IBT is used in semi-supervised MT, where a fixed model p_θ is used to back-translate the entire monolingual corpus to train q_ϕ until convergence and vice-versa for p_θ (Zhang et al., 2018).

Summary This theoretical analysis suggests that the dual reconstruction objective is a good alternative to the intractable marginal likelihood objective, and that IBT approximates it more closely than the more complex Dual Learning objective. However, we do not know whether the Dual Reconstruction optimum is reached in practice. We therefore conduct an extensive empirical study to determine whether the differences in approximations made by IBT and Dual Learning matter.

4 Empirical Study

We evaluate on six translation tasks (Table 1), including German↔English (de-en), Turkish↔English (tr-en) from WMT18 (Bojar et al., 2018), and a cross-domain task which tests de-en models trained on WMT data on the TED test sets from IWSLT17 (Cettolo et al., 2017).³

4.1 Model and Training Configuration

We adopt the base Transformer model (Vaswani et al., 2017). We pre-train models with the supervised objective until convergence, and fine-tune on the mixed parallel and monolingual data as in prior work (Sennrich et al., 2016a; Cotterell

³We exclude Rapid and ParaCrawl corpora as they are noisy and thus require data filtering (Morishita et al., 2018).

Task	Lang.	Parallel Data	Mono. Data	Validation	Test
high-resource	de-en	News 4.5M	News 5.0M	newstest15	newstest16-18
low-resource	tr-en	News 0.2M	News 0.8M	newstest16	newstest17-18
cross-domain	de-en	News 4.5M	TED 0.5M	iwslt-test14	iwslt-test15-17

Table 1: The empirical comparison spans three data conditions (and both translation directions). We report provenance and the number of sentences in parallel and monolingual training data, as well as validation and test sets for each setting. Monolingual data are randomly selected from “*News Crawl: articles from 2015*” for German↔English and “*News Crawl: articles from 2017*” for Turkish↔English, and TED talks data for TED.

Low-Resource	α_{LM}	hours	tr-en BLEU			en-tr BLEU				
			2017	2018	Avg	2017	2018	Avg		
baseline	–	8.0	15.14	15.95	15.55	11.17	10.18	10.68		
epoch-level IBT-1	–	86.1	16.36	16.44	16.40	15.08	12.98	14.03		
epoch-level IBT-2	–	162.2	19.12	19.63	19.38	14.94	12.53	13.74		
epoch-level IBT-3	–	237.5	18.76	19.01	18.89	15.04	12.93	13.99		
batch-level IBT	–	160.6	17.18	18.08	17.63	13.90	11.84	12.87		
Dual Learning	0.0	313.2	17.07	18.00	17.54	14.17	11.91	13.04		
Dual Learning	0.1	257.8	17.09	17.62	17.36	13.88	11.49	12.69		
Dual Learning	0.5	421.2	17.33	18.36	17.85	14.54	12.30	13.42		
High-Resource	α_{LM}	hours	de-en BLEU			en-de BLEU				
			2016	2017	2018	Avg	2016	2017	2018	Avg
baseline	–	26.7	31.95	27.74	34.59	31.43	29.18	23.46	34.53	29.06
epoch-level IBT-1	–	439.0	32.59	28.46	35.22	32.09	30.13	23.87	35.35	29.78
epoch-level IBT-2	–	850.9	33.64	29.13	36.37	33.05	29.99	24.42	35.60	30.00
epoch-level IBT-3	–	1261.6	33.43	29.07	36.17	32.89	29.93	24.24	35.46	29.88
batch-level IBT	–	94.0	32.95	28.65	35.24	32.28	29.70	23.78	34.89	29.46
Dual Learning	0.0	128.2	32.79	28.47	35.10	32.12	29.37	23.50	34.67	29.18
Dual Learning	0.1	93.3	32.63	28.47	34.88	31.99	29.38	23.79	34.71	29.29
Dual Learning	0.5	152.1	32.89	28.69	35.32	32.30	29.58	23.65	34.88	29.37
Cross-Domain	α_{LM}	hours	de-en BLEU			en-de BLEU				
			2015	2016	2017	Avg	2015	2016	2017	Avg
baseline	–	26.2	27.11	27.37	23.65	26.04	26.35	23.10	21.69	23.71
epoch-level IBT-1	–	71.1	28.88	28.73	25.37	27.66	26.69	24.02	22.59	24.43
epoch-level IBT-2	–	115.0	28.70	28.72	25.37	27.60	27.57	24.50	22.78	24.95
epoch-level IBT-3	–	159.8	29.13	29.00	25.33	27.82	27.31	24.37	22.92	24.87
batch-level IBT	–	45.0	28.03	27.78	24.53	26.78	26.84	23.64	22.35	24.28
Dual Learning	0.0	65.8	28.04	27.73	24.36	26.71	26.70	23.85	22.21	24.25
Dual Learning	0.1	59.3	27.77	27.84	24.51	26.71	26.99	23.86	22.59	24.48
Dual Learning	0.5	92.7	27.84	28.00	24.18	26.67	27.23	24.08	22.72	24.68

Table 2: BLEU scores and total training time (*hours*) on the low-resource, high-resource, and cross-domain tasks. *epoch-level* IBT-1, IBT-2, and IBR-3 denotes models fine-tuned with IBT for 1–3 iterations, and α_{LM} denotes the weight for the LM loss. We boldface the highest average scores and their ties based on the significance test. Overall, epoch-level IBT outperforms all other methods at the cost of much longer training time.

and Kreutzer, 2018). We use the Adam optimizer (Kingma and Ba, 2015) with a batch size of 32 sentences and checkpoint the model every 2500 updates. At decoding time, we use beam search with a beam size of 5. The LMs in Dual Learning are RNNs (Mikolov et al., 2010) with 512 hidden units. All model and training details are in Appendix B.

For preprocessing, we normalize punctuations and apply tokenization, true-casing, and joint source-target Byte Pair Encoding (Sennrich et al., 2016b) with 32,000 operations. We set the maximum sentence length to 50.

4.2 Baselines and Evaluation

Our experiments are based on strong supervised baselines.⁴ We compare semi-supervised models that are fine-tuned with Back-Translation, epoch-level and batch-level IBT, and Dual Learning with varying interpolation weights $\alpha_{LM} = \{0, 0.1, 0.5\}$ for the LM loss.⁵ Following He et al. (2016), we use beam search with a beam size of 2 for inference in Dual Learning and IBT.

We evaluate translation quality using sacre-BLEU⁶ and total training time in hours. We also show learning curves for the approximated dual reconstruction loss (negative of the dual reconstruction objective in Eq. (2), averaged over the training batches from both directions).

4.3 Findings

Overview All semi-supervised training techniques improve translation quality over the supervised-only baseline (Table 2). The first iteration of IBT (i.e. Back-Translation) on monolingual data improves over the baseline significantly⁷ by 0.7–3.4 BLEU. IBT is more effective in the direction where the model in the opposite direction is most improved by Back-Translation. For example, in the high and low resource tasks where Back-Translation improves over the baseline more when translating out of English, the best performing IBT model significantly improves

⁴de-en: 2–4 BLEU higher than the baseline of Morishita et al. (2018); tr-en: on par or higher than the baseline of García-Martínez et al. (2017).

⁵By contrast, prior work only reports results for $\alpha_{LM} = 0.005$ (He et al., 2016). Our preliminary result show that $\alpha_{LM} = 0.005$ obtains similar results to $\alpha_{LM} = 0$.

⁶Version: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.2.11

⁷All mentions of statistical significance are based on a paired Student’s t-test with $p < 0.05$.

BLEU over Back-Translation when translating into English, but not in the other direction. In the cross-domain scenario where Back-Translation improves more on de-en, IBT outperforms Back-Translation on en-de, but the improvement is not significant in the other direction.

Impact of Policy Gradient Updating the inference model via policy gradient fails to lower the dual reconstruction loss and has little impact on BLEU. We compare Dual Learning (with $\alpha_{LM} = 0$) to batch-level IBT, so that the only difference between the two approaches is whether the inference model is updated. Batch-level IBT achieves similar or higher BLEU than Dual Learning for all tasks, except for the low-resource en-tr task where the BLEU difference is small (< 0.2). In addition, batch-level IBT trains 30–50% faster than Dual Learning. Figure 2 shows that the policy gradient update has little impact on the dual reconstruction loss on all tasks.

Impact of LM The best Dual Learning BLEU is obtained with $\alpha_{LM} = 0.5$ on all tasks except for de-en in the cross-domain setting (Table 2). However, it brings only small BLEU improvements (0.2–0.4) over Dual Learning without LM loss ($\alpha_{LM} > 0$), but causes the dual reconstruction loss to decrease slower (Figure 2), and slows down training by 20–40%. In all cases, IBT outperforms Dual Learning.

Epoch vs. Batch IBT The best epoch-level IBT model outperforms batch-level IBT by 0.5–1.8 BLEU overall, at the cost of much slower training: 13 times longer in the high-resource setting, 1.5 times longer in the low-resource setting, and 3.5 times longer in the cross-domain setting. Running IBT for two iterations is a good choice to balance training efficiency and translation quality, as the third iteration does not help BLEU.

4.4 Mutual Information Analysis

We test the hypothesis that the mutual information constraint is met when training models on the combined supervised and unsupervised objectives in the low-resource setting (the most adversarial condition with the fewest supervised training samples).

The mutual information I_{p_θ} from Definition 3.2 can be computed by Hoffman and Johnson (2016):

$$I_{p_\theta} = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} [D_{\text{KL}} [p_\theta(\mathbf{y} | \mathbf{x}) || p(\mathbf{y})]] - D_{\text{KL}} [p_\theta(\mathbf{y}) || p(\mathbf{y})] \quad (8)$$

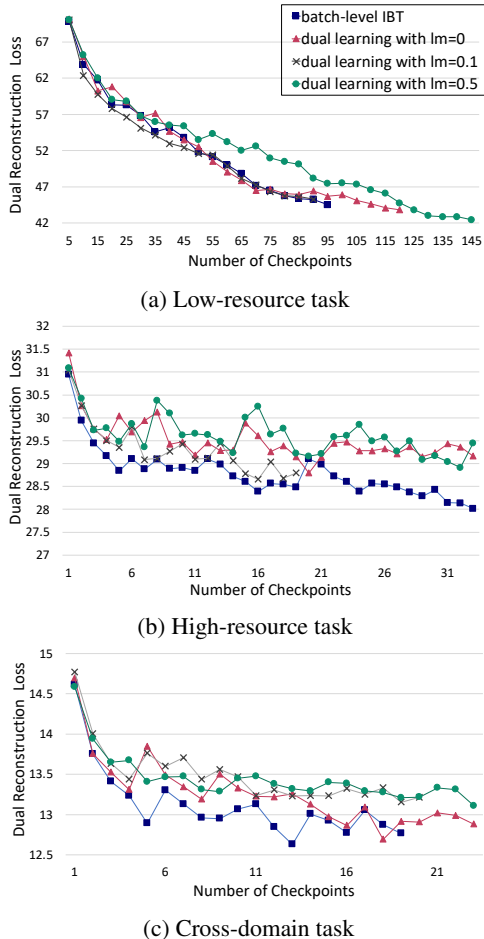


Figure 2: Learning curves for the approximated dual reconstruction loss averaged over the training batches from both directions on the low-resource, high-resource, and cross-domain tasks.

where prior distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ are estimated by the empirical data distribution given the monolingual corpora \mathcal{M}_X and \mathcal{M}_Y . Although computing I_{p_θ} directly is intractable, it can be approximated with a Monte Carlo estimate. Following [Dieng et al. \(2019\)](#), we approximate the two KL terms by Monte Carlo, where samples from $p_\theta(\mathbf{y})$ can be obtained by ancestral sampling (we use beam search with beam size of five to sample from $p_\theta(\mathbf{y}|\mathbf{x})$). The marginal probability $p_\theta(\mathbf{y}) = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} [p_\theta(\mathbf{y}|\mathbf{x})]$ can also be estimated by Monte Carlo. Due to data sparsity, the conditional likelihood $p_\theta(\mathbf{y}|\mathbf{x})$ will be near zero for most source sentences randomly sampled from $q(\mathbf{x})$. To better estimate it, we smooth the data distribution of the original dataset \mathcal{D} by generating a randomly perturbed dataset $\tilde{\mathcal{D}}$.⁸

⁸We generate 20 perturbed sentences per source via random word dropping with probability of 0.1 and permutation with maximum distance of 3.

	tr-en	en-tr
baseline	-2.47	-2.28
epoch-level IBT-1	-2.57	-2.23
epoch-level IBT-2	-2.18	-2.30
epoch-level IBT-3	-2.32	-2.42
batch-level IBT	-1.51	-1.82
dual learning w/ $\alpha_{LM} = 0$	-1.50	-1.80
dual learning w/ $\alpha_{LM} = 0.5$	-1.44	-1.77

Table 3: Results on estimated mutual information \tilde{I} in the low-resource setting. We report the normalized scores $\tilde{I} - \log |\mathcal{D}|$ (on the scale of 10^{-4}) averaged over the two test sets. The range of normalized scores should be $[-\log |\mathcal{D}|, \log \frac{|\tilde{\mathcal{D}}|}{|\mathcal{D}|}] = [-8.0, 3.0]$.

Table 3 shows the normalized mutual information $\tilde{I} - \log |\mathcal{D}|$ where \tilde{I} denotes the estimated mutual information. It shows that, when training with the combination of supervised and unsupervised objectives, the normalized mutual information is within a small range between $(-2.6 \times 10^{-4}, -1.4 \times 10^{-4})$ and is lower than the maximum normalized mutual information $\log |\tilde{\mathcal{D}}| - \log |\mathcal{D}| \approx 3.0$ by a large margin. Thus, the mutual information can be bounded by appropriate values of I_{min} and I_{max} to satisfy the constraint. In addition, these results confirm that updating the inference model using policy gradient in Dual Learning does not effectively increase model’s mutual information.

5 Summary of Contributions

We contribute theoretical and empirical results that improve our understanding of the connection between two seemingly distant semi-supervised training strategies for NMT: Iterative Back-Translation (IBT) and Dual Learning.

On the theory side, we define a dual reconstruction objective which unifies semi-supervised NMT techniques that exploit source and target monolingual text. We prove that optimizing this objective leads to the same global optimum as the intractable marginal likelihood objective, where the model’s marginal distribution coincides with the prior language distribution while also maximizing the model’s mutual information between source and target. IBT approximates this objective more closely than Dual Learning, despite the more complex objective and update strategies used in the latter.

We present a systematic empirical comparison of Back-Translation, IBT, and Dual Learning on six

tasks spanning high-resource, low-resource, and cross-domain settings. Results support the theory that the LM loss and policy gradient estimation are unnecessary in Dual Learning, and show that IBT achieves better translation quality than Dual Learning. Analysis confirms that the mutual information constraint required to reach an interesting dual reconstruction optimum is satisfied in practice.

These findings lead us to recommend batch-level IBT to quickly boost model performance at early training stages and epoch-level IBT to further improve quality. Our theory also suggests future directions for improving unsupervised MT via more effective methods to maximize the model’s mutual information between source and target, and the potential of applying our dual reconstruction objective to other sequence-to-sequence tasks.

Acknowledgments

We thank the anonymous reviewers, Jordan Boyd-Graber, Hal Daumé III, Naomi Feldman, Shi Feng, Pranav Goel, Alexander Miserlis Hoyle, Michelle Yuan, Mozhi Zhang, and the CLIP lab at UMD for helpful comments. This research is supported in part by an Amazon Web Services Machine Learning Research Award and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Mikel Artetxe, Gorika Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. [Unsupervised neural machine translation](#). In *Proceedings of the 6th International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. [Neural machine translation by jointly learning to align and translate](#). In *Proceedings of the 3th International Conference on Learning Representations*.

Christos Baziotis, Ion Androustopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. [SEQ³: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence](#)

[compression](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

Ondrej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. [Findings of the 2018 conference on machine translation \(WMT18\)](#). In *Proceedings of the Third Conference on Machine Translation*, pages 272–307. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Nihues Jan, Stüker Sebastian, Sudoh Katsuiho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the IWSLT 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Ryan Cotterell and Julia Kreutzer. 2018. [Explaining and generalizing back-translation through wake-sleep](#). *CoRR*, abs/1806.04402.

Adji B Dieng, Yoon Kim, Alexander M Rush, and David M Blei. 2019. Avoiding latent variable collapse with generative skip models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2397–2405.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. [Understanding back-translation at scale](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. Association for Computational Linguistics.

Mercedes García-Martínez, Ozan Caglayan, Walid Aransa, Adrien Bardet, Fethi Bougares, and Loïc Barrault. 2017. [LIUM machine translation systems for WMT17 news translation task](#). In *Proceedings of the Second Conference on Machine Translation*, pages 288–295. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. [Achieving human parity on automatic Chinese to English news translation](#). *CoRR*, abs/1803.05567.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. [Dual learning for machine translation](#). In *Advances in Neural Information Processing Systems 29*, pages 820–828. Curran Associates, Inc.

- Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. [Deep reinforcement learning that matters](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press.
- M. Hledík, T. R. Sokolowski, and G. Tkačik. 2019. A tight upper bound on mutual information. In *2019 IEEE Information Theory Workshop (ITW)*, pages 1–5.
- Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. [Iterative back-translation for neural machine translation](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24. Association for Computational Linguistics.
- Matthew D Hoffman and Matthew J Johnson. 2016. [ELBO surgery: yet another way to carve up the variational evidence lower bound](#). In *Proceedings of NIPS Workshop in Advances in Approximate Bayesian Inference*.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with gumbel-softmax](#). In *Proceedings of the 5th International Conference on Learning Representations*.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *Proceedings of the 3th International Conference on Learning Representations*.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. [Unsupervised machine translation using monolingual corpora only](#). In *Proceedings of the 6th International Conference on Learning Representations*.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. [Phrase-based & neural unsupervised machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049. Association for Computational Linguistics.
- Yishu Miao and Phil Blunsom. 2016. [Language as a latent variable: Discrete generative models for sentence compression](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 319–328, Austin, Texas. Association for Computational Linguistics.
- Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.
- Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2018. [NTT’s neural machine translation systems for WMT 2018](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 461–466, Belgium, Brussels. Association for Computational Linguistics.
- Toan Q. Nguyen and David Chiang. 2018. [Improving lexical choice in neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 334–343. Association for Computational Linguistics.
- Xing Niu, Michael Denkowski, and Marine Carpuat. 2018. [Bi-directional neural machine translation with synthetic parallel data](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 84–91. Association for Computational Linguistics.
- Ofir Press and Lior Wolf. 2017. [Using the output embedding to improve language models](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Computational*, pages 157–163. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.
- Lierni Sestorain, Massimiliano Ciaramita, Christian Buck, and Thomas Hofmann. 2018. [Zero-shot dual machine translation](#). *arXiv preprint arXiv:1805.10338*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2019. [Multi-agent dual learning](#). In *International Conference on Learning Representations*.
- Ronald J. Williams. 1992. [Simple statistical gradient-following algorithms for connectionist reinforcement learning](#). *Machine Learning*, 8(3):229–256.
- Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. [A study of reinforcement learning for neural machine translation](#). In *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3612–3621. Association for Computational Linguistics.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. [Joint training for neural machine translation models with monolingual data](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 555–562. AAAI Press.

A Proof

A.1 Proof for Proposition 1

Proposition 1. Given prior distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ over $\mathbf{x} \in \Sigma_x$ and $\mathbf{y} \in \Sigma_y$, if parameterized probability models p_θ and q_ϕ have enough capacity under the constraint that:

$$0 \leq I_{min} \leq I_{p_\theta}, I_{q_\phi} \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y})$$

where I_{min} and I_{max} are pre-defined constant values between zero and the maximum mutual information between \mathbf{x} and \mathbf{y} given any joint distribution $p(\mathbf{x}, \mathbf{y}) \in P_{XY}$ whose marginals satisfy $\sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})$ and $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})$. Then, the dual reconstruction objective is upper-bounded by $\mathcal{J}_{dual}(\theta, \phi) \leq 2I_{max} - \mathbb{H}[q(\mathbf{x})] - \mathbb{H}[p(\mathbf{y})]$, and the upper bound is achieved iff

$$\begin{aligned} I_{q_\phi} &= I_{max} \\ I_{p_\theta} &= I_{max} \\ p_\theta(\mathbf{y} | \mathbf{x}) &= \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y}) \\ q_\phi(\mathbf{x} | \mathbf{y}) &= \frac{p_\theta(\mathbf{y} | \mathbf{x})}{p_\theta(\mathbf{y})} q(\mathbf{x}) \end{aligned} \quad (9)$$

Proof. First we prove that $\mathcal{J}_1(\theta, \phi) \leq I_{max} - \mathbb{H}[p(\mathbf{y})]$, and the upper bound is achieved iff

$$\begin{aligned} I_{q_\phi} &= I_{max} \\ p_\theta(\mathbf{y} | \mathbf{x}) &= \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y}) \end{aligned}$$

where $\mathbb{H}[p(\mathbf{y})]$ is the entropy of the prior distribution $p(\mathbf{y})$.

To show this, we denote the posterior distribu-

tion $Q(\mathbf{y} | \mathbf{x}) = \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y})$, and rewrite \mathcal{J}_1 :

$$\begin{aligned} \mathcal{J}_1 &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} [\log p_\theta(\mathbf{y} | \mathbf{x})] \right] \\ &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} [\log Q(\mathbf{y} | \mathbf{x})] \right] \\ &\quad + \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} \left[\log \frac{p_\theta(\mathbf{y} | \mathbf{x})}{Q(\mathbf{y} | \mathbf{x})} \right] \right] \\ &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} [\log Q(\mathbf{y} | \mathbf{x})] \right] \\ &\quad + \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} \left[\log \frac{p_\theta(\mathbf{y} | \mathbf{x}) q_\phi(\mathbf{x})}{q_\phi(\mathbf{x} | \mathbf{y}) p(\mathbf{y})} \right] \right] \\ &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} [\log Q(\mathbf{y} | \mathbf{x})] \right] \\ &\quad - D_{KL} [q_\phi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) || p_\theta(\mathbf{y} | \mathbf{x}) q_\phi(\mathbf{x})] \\ &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} \left[\log \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} \right] \right] \\ &\quad + \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_\phi(\mathbf{x} | \mathbf{y})} [\log p(\mathbf{y})] \right] \\ &\quad - D_{KL} [q_\phi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) || p_\theta(\mathbf{y} | \mathbf{x}) q_\phi(\mathbf{x})] \\ &= I_{q_\phi} - \mathbb{H}[p(\mathbf{y})] \\ &\quad - D_{KL} [q_\phi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) || p_\theta(\mathbf{y} | \mathbf{x}) q_\phi(\mathbf{x})] \end{aligned}$$

Since the KL divergence between two distributions is always non-negative and is zero iff they are equal, we have

$$\mathcal{J}_1(\theta, \phi) \leq I_{q_\phi} - \mathbb{H}[p(\mathbf{y})] \leq I_{max} - \mathbb{H}[p(\mathbf{y})]$$

and $\mathcal{J}_1(\theta, \phi) = I_{max} - \mathbb{H}[p(\mathbf{y})]$ iff

$$\begin{aligned} I_{q_\phi} &= I_{max} \\ D_{KL} [q_\phi(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) || p_\theta(\mathbf{y} | \mathbf{x}) q_\phi(\mathbf{x})] &= 0 \end{aligned}$$

The second equality holds iff

$$p_\theta(\mathbf{y} | \mathbf{x}) = \frac{q_\phi(\mathbf{x} | \mathbf{y})}{q_\phi(\mathbf{x})} p(\mathbf{y})$$

Similarly, we can prove that $\mathcal{J}_2(\theta, \phi) \leq I_{max} - \mathbb{H}[q(\mathbf{x})]$, and the upper bound is achieved iff

$$\begin{aligned} I_{p_\theta} &= I_{max} \\ q_\phi(\mathbf{x} | \mathbf{y}) &= \frac{p_\theta(\mathbf{y} | \mathbf{x})}{p_\theta(\mathbf{y})} q(\mathbf{x}) \end{aligned}$$

thus $\mathcal{J}_{dual}(\theta, \phi) \leq 2I_{max} - \mathbb{H}[q(\mathbf{x})] - \mathbb{H}[p(\mathbf{y})]$ and the upper bound is achieved iff θ and ϕ satisfy Eq. (9), concluding the proof. \square

A.2 Proof for Proposition 2

Proposition 2. Given distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ over $\mathbf{x} \in \Sigma_x$ and $\mathbf{y} \in \Sigma_y$, if parameterized probability models p_θ and q_ϕ have enough capacity under

the constraint that:

$$0 \leq I_{min} \leq I_{p_\theta}, I_{q_\phi} \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y}) \quad (10)$$

where I_{min} and I_{max} are pre-defined constant values between zero and the maximum mutual information between \mathbf{x} and \mathbf{y} given any joint distribution $p(\mathbf{x}, \mathbf{y}) \in P_{XY}$ whose marginals satisfy $\sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})$ and $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})$. Then there exist θ^* and ϕ^* such that:

$$\begin{aligned} I_{q_{\phi^*}} &= I_{p_{\theta^*}} = I_{max} \\ p_{\theta^*}(\mathbf{y} | \mathbf{x}) &= \frac{q_{\phi^*}(\mathbf{x} | \mathbf{y})}{q_{\phi^*}(\mathbf{x})} p(\mathbf{y}) \\ q_{\phi^*}(\mathbf{x} | \mathbf{y}) &= \frac{p_{\theta^*}(\mathbf{y} | \mathbf{x})}{p_{\theta^*}(\mathbf{y})} q(\mathbf{x}) \end{aligned} \quad (11)$$

Proof. Since I_{max} satisfies

$$0 = \min_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y}) \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y})$$

there exists a joint distribution $p^*(\mathbf{x}, \mathbf{y}) \in P_{XY}$ such that

$$I_{p^*}(\mathbf{x}; \mathbf{y}) = I_{max}$$

As models p_θ and q_ϕ have enough capacity under the constraint in Eq. (10), there exist θ^* and ϕ^* such that $\forall \mathbf{x} \in \Sigma_x, \forall \mathbf{y} \in \Sigma_y$

$$\begin{aligned} p_{\theta^*}(\mathbf{y} | \mathbf{x}) &= \frac{p^*(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})} \\ q_{\phi^*}(\mathbf{x} | \mathbf{y}) &= \frac{p^*(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} \end{aligned}$$

thus

$$\begin{aligned} p_{\theta^*}(\mathbf{y}) &= \sum_{\mathbf{x}} p_{\theta^*}(\mathbf{y} | \mathbf{x}) q(\mathbf{x}) = \sum_{\mathbf{x}} p^*(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}) \\ q_{\phi^*}(\mathbf{x}) &= \sum_{\mathbf{y}} q_{\phi^*}(\mathbf{x} | \mathbf{y}) p(\mathbf{y}) = \sum_{\mathbf{y}} p^*(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}) \end{aligned}$$

and thus

$$\begin{aligned} I_{p_{\theta^*}} &= \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \left[\mathbb{E}_{\mathbf{y} \sim p_{\theta^*}(\mathbf{y} | \mathbf{x})} \left[\log \frac{p_{\theta^*}(\mathbf{y} | \mathbf{x})}{p_{\theta^*}(\mathbf{y})} \right] \right] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p^*(\mathbf{x}, \mathbf{y})} \left[\log \frac{p^*(\mathbf{x}, \mathbf{y})}{q(\mathbf{x}) p(\mathbf{y})} \right] \\ &= I_{p^*}(\mathbf{x}; \mathbf{y}) \\ &= I_{max} \end{aligned}$$

$$\begin{aligned} I_{q_{\phi^*}} &= \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} \left[\mathbb{E}_{\mathbf{x} \sim q_{\phi^*}(\mathbf{x} | \mathbf{y})} \left[\log \frac{q_{\phi^*}(\mathbf{x} | \mathbf{y})}{q_{\phi^*}(\mathbf{x})} \right] \right] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim p^*(\mathbf{x}, \mathbf{y})} \left[\log \frac{p^*(\mathbf{x}, \mathbf{y})}{q(\mathbf{x}) p(\mathbf{y})} \right] \\ &= I_{p^*}(\mathbf{x}; \mathbf{y}) \\ &= I_{max} \end{aligned}$$

and

$$\frac{q_{\phi^*}(\mathbf{x} | \mathbf{y})}{q_{\phi^*}(\mathbf{x})} p(\mathbf{y}) = \frac{p^*(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} \frac{p(\mathbf{y})}{q(\mathbf{x})} = p_{\theta^*}(\mathbf{y} | \mathbf{x})$$

$$\frac{p_{\theta^*}(\mathbf{y} | \mathbf{x})}{p_{\theta^*}(\mathbf{y})} q(\mathbf{x}) = \frac{p^*(\mathbf{x}, \mathbf{y})}{q(\mathbf{x})} \frac{q(\mathbf{x})}{p(\mathbf{y})} = q_{\phi^*}(\mathbf{x} | \mathbf{y})$$

concluding the proof. \square

A.3 Proof for Theorem 1

Theorem 1. Given prior distributions $q(\mathbf{x})$ and $p(\mathbf{y})$ over $\mathbf{x} \in \Sigma_x$ and $\mathbf{y} \in \Sigma_y$, if parameterized probability models p_θ and q_ϕ have enough capacity under the constraint that:

$$0 \leq I_{min} \leq I_{p_\theta}, I_{q_\phi} \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y})$$

where I_{min} and I_{max} are pre-defined constant values between zero and the maximum mutual information between \mathbf{x} and \mathbf{y} given any joint distribution $p(\mathbf{x}, \mathbf{y}) \in P_{XY}$ whose marginals satisfy $\sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y})$ and $\sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})$. Let θ^*, ϕ^* be the global optimum of the dual reconstruction objective $\max_{\theta, \phi} \mathcal{J}_{dual}(\theta, \phi)$, then $q_{\phi^*}(\mathbf{x}) = q(\mathbf{x})$, $p_{\theta^*}(\mathbf{y}) = p(\mathbf{y})$, and $I_{q_{\phi^*}} = I_{p_{\theta^*}} = I_{max}$.

Proof. Suppose models p_θ and q_ϕ have enough capacity under the constraint that:

$$0 \leq I_{min} \leq I_{p_\theta}, I_{q_\phi} \leq I_{max} \leq \max_{p \in P_{XY}} I_p(\mathbf{x}; \mathbf{y})$$

then based on Proposition 1, $\mathcal{J}_{dual}(\theta, \phi) \leq 2I_{max} - \mathbb{H}[q(\mathbf{x})] - \mathbb{H}[p(\mathbf{y})]$, and the upper bound is achieved iff the optimal criteria Eq. (9) hold. And based on Proposition 2, there exists a solution p_{θ^*} and q_{ϕ^*} for the criteria Eq. (9). Thus

$$\max_{\theta, \phi} \mathcal{J}_{dual}(\theta, \phi) = 2I_{max} - \mathbb{H}[q(\mathbf{x})] - \mathbb{H}[p(\mathbf{y})]$$

Based on the first equation in Eq. (9), we have:

$$I_{q_{\phi^*}} = I_{p_{\theta^*}} = I_{max} \quad (12)$$

And multiply the last two equations, we have:

$$p(\mathbf{y}) q(\mathbf{x}) = q_{\phi^*}(\mathbf{x}) p_{\theta^*}(\mathbf{y}) \quad (13)$$

Given Lemma 1, we have $q_{\phi^*}(\mathbf{x}) = q(\mathbf{x})$ and $p_{\theta^*}(\mathbf{y}) = p(\mathbf{y})$, concluding the proof. \square

Task	NMT.xx-en	NMT.en-xx	LM.xx	LM.en
low-resource	22.12	30.92	121.13	78.96
high-resource	6.72	6.25	78.32	74.00
cross-domain	8.17	7.35	102.43	92.70

Table 4: Validation perplexity of the NMT and LM models. We denote English as *en* and the other language as *xx*.

Task	NMT.xx-en	NMT.en-xx	LM.xx	LM.en
low-resource	92811565	92811565	27311123	19247448
high-resource	98346302	98346302	32165523	24095698
cross-domain	98346302	98346302	24806023	18768773

Table 5: Number of model parameters. We denote English as *en* and the other language as *xx*.

Lemma 1. Let $p(x)$ and $p'(x)$ be two discrete probability functions over random variable $x \in \Sigma_x$, and $q(y)$ and $q'(y)$ be two discrete probability functions over random variable $y \in \Sigma_y$.

If $\forall x \in \Sigma_x, \forall y \in \Sigma_y, p'(x)q'(y) = p(x)q(y)$, then $\forall x \in \Sigma_x, \forall y \in \Sigma_y, p'(x) = p(x)$ and $q'(y) = q(y)$.

Proof. Let $x_0 \in \Sigma_x$ such that $p(x_0) \neq 0$ and $p'(x_0) \neq 0$, and $y_0 \in \Sigma_y$ such that $q(y_0) \neq 0$ and $q'(y_0) \neq 0$.

Since

$$p'(x_0)q'(y_0) = p(x_0)q(y_0) \quad (14)$$

and for any $x \in \Sigma_x$

$$p'(x)q'(y_0) = p(x)q(y_0) \quad (15)$$

we have

$$\frac{p'(x)}{p'(x_0)} = \frac{p(x)}{p(x_0)} \quad (16)$$

Given $\sum_{x \in \Sigma_x} p'(x) = 1$, $\sum_{x \in \Sigma_x} p(x) = 1$, and since

$$\sum_{x \in \Sigma_x} p'(x) = \frac{p'(x_0)}{p(x_0)} \sum_{x \in \Sigma_x} p(x) \quad (17)$$

we have $p'(x_0) = p(x_0)$.

Thus for any $x \in \Sigma_x$, $p'(x) = \frac{p'(x_0)}{p(x_0)} p(x) = p(x)$. For any $y \in \Sigma_y$, $q'(y) = \frac{p(x)}{p'(x)} q(y) = q(y)$, concluding the proof. \square

B Experimental Setup

B.1 Tasks and Data

We evaluate on six translation tasks including German \leftrightarrow English (de-en),⁹ Turkish \leftrightarrow English (tr-

⁹We exclude Rapid and ParaCrawl corpora as they are noisy and thus require data filtering (Morishita et al., 2018).

en) from WMT18 (Bojar et al., 2018),¹⁰ and a cross-domain task which tests de \leftrightarrow en models trained on WMT data on the TED test sets from IWSLT17 (Cettolo et al., 2017).¹¹

B.2 Model and Training Configuration

We adopt the base Transformer model (Vaswani et al., 2017) with $d_{\text{model}} = 512$, $d_{\text{hidden}} = 2048$, $n_{\text{heads}} = 8$, $m_{\text{layers}} = 6$, and $p_{\text{drop}} = 0.1$. We tie the source and target embeddings with the output layer weights (Press and Wolf, 2017; Nguyen and Chiang, 2018).

We use the Adam optimizer (Kingma and Ba, 2015) with a batch size of 32 sentences and checkpoint the model every 2500 updates. Training hyperparameters and stopping criteria are constant across all comparable experimental conditions. Initial learning rates for pre-training and fine-tuning are respectively set to 10^{-4} and 2×10^{-5} . We decay the learning rate by 30% and reload the best model after 3 checkpoints without improvement. We apply early stopping after repeating this process for 5 times. We adopt the same learning rate decay and stopping criteria during fine-tuning. For batch-level IBT and Dual Learning, we check whether both models improve validation perplexity. For epoch-level IBT, we run for 3 iterations.

The LMs in Dual Learning are RNNs (Mikolov et al., 2010) with 512 hidden units, embeddings of size 512, and dropout of 0.2 to hidden states. We tie the input embeddings with the output layer weights. We clip the gradients at a threshold of 5. We train them similarly to NMT models, except setting the batch size to 64 sentences and the initial

¹⁰<http://www.statmt.org/wmt18/translation-task.html>

¹¹<https://wit3.fbk.eu/mt.php?release=2017-01-ted-test>

learning rate to 0.001. We decay the learning rate by 50% and reload the best model after 5 checkpoints without validation perplexity improvement and apply early stopping after repeating the process for 5 times. We report the validation perplexity of the NMT and LM models in Table 4, and the model sizes in Table 5. All experiments are performed on a single NVIDIA GeForce GTX 1080 Ti GPU.