Uncertainty-Aware Semantic Augmentation for Neural Machine Translation

Xiangpeng Wei[1,2][*], **Heng Yu**[3], **Yue Hu**[1,2][†], **Rongxiang Weng**[3], **Luxi Xing**[1,2], **Weihua Luo**[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Machine Intelligence Technology Lab, Alibaba Group, Hangzhou, China
{weixiangpeng,huyue,xingluxi}@iie.ac.cn
{yuheng.yh,wengrx,weihua.luowh}@alibaba-inc.com

Abstract

As a sequence-to-sequence generation task, neural machine translation (NMT) naturally contains *intrinsic uncertainty*, where a single sentence in one language has multiple valid counterparts in the other. However, the dominant methods for NMT only observe one of them from the parallel corpora for the model training but have to deal with adequate variations under the same meaning at inference. This leads to a discrepancy of the data distribution between the training and the inference phases. To address this problem, we propose *uncertainty-aware semantic augmentation*, which explicitly captures the universal semantic information among multiple semantically-equivalent source sentences and enhances the hidden representations with this information for better translations. Extensive experiments on various translation tasks reveal that our approach significantly outperforms the strong baselines and the existing methods.

1 Introduction

In recent years neural machine translation (NMT) has demonstrated state-of-the-art performance on many language pairs with advanced architectures and large scale data (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017). At training time the parallel data only contains one source sentence as the input and the rest reasonable ones are ignored, while at inference the resulting model has to deal with adequate variations under the same meaning. This discrepancy of the data distribution poses a formidable learning challenge of the inherent uncertainty in machine translation. Since typically there are several semantically-equivalent source sentences that can be translated to the same target sentence, but the model only observes one at

training time. Thus it is natural to enable an NMT model trained with the token-level cross-entropy (CE) to capture such a rich distribution, which exactly motivates our work.

Intuitively, the NMT model should be trained under the guidance of the same latent semantics that it will access at inference time. In their seminal work, the variational models (Blunsom et al., 2008; Zhang et al., 2016; Shah and Barber, 2018) introduce a continuous latent variable to serve as a global semantic signal to guide the generation of target translations. Wei et al. (2019) consider an universal topic representation for each sentence pair as global semantics for enhancing representations learnt by NMT models. Yang et al. (2019) minimize the difference between the representation of source and target sentences. Although their methods yield notable results, they are still limited to one-to-one parallel sentence pairs.

To address these problems, we present a novel *uncertainty-aware semantic augmentation* method, which takes account of the intrinsic uncertainty sourced from the one-to-many nature of machine translation (Ott et al., 2018). Specifically, we first synthesize multiple reasonable source sentences to play the role of inherent uncertainty[1] for each target sentence. To achieve this, we introduce a *controllable sampling* strategy to cover adequate variations for inputs, by quantifying the sharpness of the word distribution in each decoding step and taking the proper word (the one with the maximum probability if sharp enough or determined by multinomial sampling) as the output. Then a *semantic constrained network* (SCN) is developed to summarize multiple source sentences that share the same meaning into a *closed semantic region*, augmented

---

[1]In our scenario, we mainly study the uncertainty in source-side, as it is problematic if the synthetic targets that inevitably contain noise and errors are used to supervise the training of models. We leave further study on this to future work.

---

[*]Work done at Alibaba Group.
[†]Corresponding Author.

by which the model generates translations finally. By integrating such *soft* correspondences into the translation process, the model can intuitively work well when fed with an unfamiliar literal expression that can be supported by its underlying semantics. In addition, given the effectiveness of leveraging monolingual data in improving translation quality (Sennrich et al., 2016a), we further propose to combine the strength of both semantic augmentation and massive monolingual data distributed in the target language.

We conduct extensive experiments in both a supervised setup with bilingual data only, and a semi-supervised setup where both bilingual and target monolingual data are available. We evaluate the proposed approach on the widely used WMT14 English→French, WMT16 English→German, NIST Chinese→English and WMT18 Chinese→English benchmarks. Experimental results show that the proposed approach consistently improves translation performance on multiple language pairs. As another bonus, by adding monolingual data in German, our approach yields an additional gain of +1.5∼+3.3 BLEU points on WMT16 English→German task. Extensive analyses reveal that:

- Our approach demonstrates strong capability on learning semantic representations.

- The proposed controllable sampling strategy introduces reasonable uncertainties into the training data and generates sentences are of both high diverse and high quality.

- Our approach motivates the models to be consistent when processing equivalent source inputs with various lteral expressions.

2 Preliminary

Neural Machine Translation (Bahdanau et al., 2015) directly models the translation probability of a target sentence $\mathbf{y} = y_1, ..., y_{T_y}$ given its corresponding source sentence $\mathbf{x} = x_1, ..., x_{T_x}$:

$$P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{T_y} P(y_i|\mathbf{y}_{<i}, \mathbf{x}; \boldsymbol{\theta}), \quad (1)$$

where $\boldsymbol{\theta}$ is a set of model parameters and $\mathbf{y}_{<i}$ is a partial translation. The word-level translation probability is formulated as: $P(y_i|\mathbf{y}_{<i}, \mathbf{x}; \boldsymbol{\theta}) \propto \exp\{g(y_{i-1}, s_i, c_i; \boldsymbol{\theta})\}$, in which $g(\cdot)$ denotes a non-linear function to predict the i-th target word y_i from the decoder state s_i and the context vector c_i summarized from a sequence of representations of the encoder with an attention module. For training, given a parallel corpus $\{(\mathbf{x}^n, \mathbf{y}^n)\}_{n=1}^{\mathcal{N}}$, the objective is to maximize $\log P(\mathbf{y}^n|\mathbf{x}^n; \boldsymbol{\theta})$ over the entire training set.

Related Work on Data augmentation. DA has been used to improve the diversity of training signals for NMT models, like randomly shuffle (swap) or drop some words in a sentence (Iyyer et al., 2015; Artetxe et al., 2018; Lample et al., 2018), randomly replace one word in the original sentences with another word (Fadaee et al., 2017; Xie et al., 2017; Kobayashi, 2018; Wang et al., 2018; Cheng et al., 2018; Gao et al., 2019), syntax-aware methods (Duan et al., 2020), as well as using target monolingual data (Sennrich et al., 2016a; Cheng et al., 2016; He et al., 2016; Zhang et al., 2018; Wu et al., 2018; Hoang et al., 2018; Niu et al., 2018; Edunov et al., 2018; Imamura et al., 2018; Xia et al., 2019). More recently, Fadaee and Monz (2018) introduce several variations of sampling strategies targeting difficult-to-predict words. Li et al. (2019) have studied that what benefits from data augmentation across different methods and tasks. Cheng et al. (2019) propose to improve the robustness of NMT models towards perturbations and minor errors by introducing adversarial inputs into training process. In contrast, we aim at bridging the discrepancy of the data distribution between the training and the inference phases, through augmenting each training instance with multiple semantically-equivalent source inputs.

Related Work on Uncertainty in NMT. Recently, there are increasing number of studies investigating the effects of quantifying uncertainties in different applications (Kendall et al., 2017; Kendall and Gal, 2017; Xiao and Wang, 2018; Zhang et al., 2019b,a; Shen et al., 2019). However, most work in NMT has focused on improving accuracy without much consideration for the intrinsic uncertainty of the translation task itself. In their seminal work, the latent variable models (Blunsom et al., 2008; Zhang et al., 2016) introduce a (set of) continuous latent variable(s) to model underlying semantics of source sentences and to guide the generation of target translations. Zaremoodi and Haffari (2018) propose a forest-to-sequence NMT model to make use of exponentially many parse trees of the source
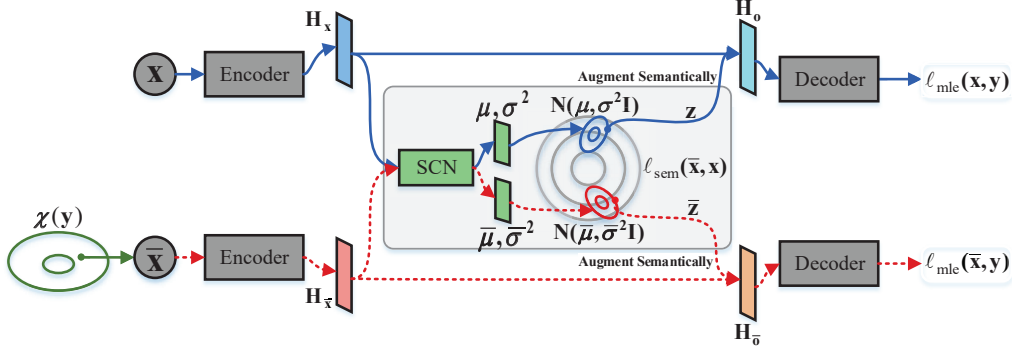
Figure 1: Uncertainty-Aware Semantic Augmentation for NMT. $\mathcal{X}(\mathbf{y})$ indicates a set of semantically-equivalent source sentences for $\mathbf{y}$. The blue-solid and red-dashed lines represent the forward-pass information flow for $\mathbf{x}$ and $\bar{\mathbf{x}}$, respectively. Note that our method involves a *shared encoder* as well as a *shared decoder* for both $\mathbf{x}$ and $\bar{\mathbf{x}}$.

sentence. Ott et al. (2018) have focused on analyzing the uncertainty in NMT that demonstrate how uncertainty is captured by the model distribution and how it affects search strategies. (Wang et al., 2019) propose to quantify the confidence of NMT model predictions based on model uncertainty. Our work significantly differs from theirs. We model the inherent uncertainty by representing multiple source sentences into a closed semantic region, and use this semantic information to enhance NMT models where diverse literal expressions intuitively be supported by their underlying semantics.

3 Uncertainty-Aware Semantic Augmentation for NMT

Here, we present the *uncertainty-aware semantic augmentation* (as shown in Figure 1), which takes account of the intrinsic uncertainty of machine translation and enhances the latent representation semantically. For each sentence-pair $(\mathbf{x}, \mathbf{y})$, supposing $\mathcal{X}(\mathbf{y})$ is a set of correct source sentences for $\mathbf{y}$, in which each sentence $\bar{\mathbf{x}}$ is assumed to have the same meaning as $\mathbf{x}$. Given a training corpus $\mathcal{D}$, we introduce the objective function as:

$$
\begin{aligned}
&\mathcal{J}(\boldsymbol{\theta}) \\
&= \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \Big\{ \lambda_1 \underbrace{\mathbb{E}_{\mathcal{P}_\phi(\mathbf{z}|\mathbf{x})}[\log P(\mathbf{y}|\mathbf{z},\mathbf{x};\boldsymbol{\theta})]}_{\ell_{\text{mle}}(\mathbf{x},\mathbf{y};\mathbf{z})} \\
&\quad - \gamma \mathbb{E}_{\bar{\mathbf{x}}\sim\mathcal{X}(\mathbf{y})}\big[\underbrace{\mathbf{KL}\big(\mathcal{P}_\phi(\bar{\mathbf{z}}|\bar{\mathbf{x}})||\mathcal{P}_\phi(\mathbf{z}|\mathbf{x})\big)}_{\ell_{\text{sem}}(\bar{\mathbf{x}},\mathbf{x})} \big] \\
&\quad + \lambda_2 \mathbb{E}_{\bar{\mathbf{x}}\sim\mathcal{X}(\mathbf{y})}\big[\underbrace{\mathbb{E}_{\mathcal{P}_\phi(\bar{\mathbf{z}}|\bar{\mathbf{x}})}[\log P(\mathbf{y}|\bar{\mathbf{z}},\bar{\mathbf{x}};\boldsymbol{\theta})]}_{\ell_{\text{mle}}(\bar{\mathbf{x}},\mathbf{y};\bar{\mathbf{z}})} \big] \Big\},
\end{aligned}
$$

(2)

where

- $\ell_{\text{sem}}(\bar{\mathbf{x}}, \mathbf{x})$ to encourage the SCN to extract the core semantics ($\bar{\mathbf{z}}$ and $\mathbf{z}$) for $\bar{\mathbf{x}}$ and $\mathbf{x}$ respectively, while constraining them into a closed semantic region. It is formulated as the negative Kullback-Leibler (**KL**) divergence between the semantic distributions $\mathcal{P}_\phi(\bar{\mathbf{z}}|\bar{\mathbf{x}})$ and $\mathcal{P}_\phi(\mathbf{z}|\mathbf{x})$, where ϕ denotes the combined parameters of the encoder and the SCN.

- $\ell_{\text{mle}}(\mathbf{x}, \mathbf{y}; \mathbf{z})$ and $\ell_{\text{mle}}(\bar{\mathbf{x}}, \mathbf{y}; \bar{\mathbf{z}})$ to guide the decoder to generate the output $\mathbf{y}$ with the assist of input-invariant semantics given diverse inputs $\mathbf{x}$ and $\bar{\mathbf{x}}$.

- λ_1 and λ_2 control the balance between the original source sentence $\mathbf{x}$ and its reasonable counterparts $\mathcal{X}(\mathbf{y})$. In experiments, we set $\lambda_1 + \lambda_2 = 1.0$, which means a target sentence occurs once in total. γ controls the impact of the semantic agreement training to be described in Section 3.3.

Intuitively, our new objective is exactly a regularized version of the widely used maximum likelihood estimation (MLE) in conventional NMT. The models are trained to optimize both the translation loss and the semantic agreement between $\mathbf{x}$ and $\bar{\mathbf{x}}$.

In the following sections, we will first describe how to summarize multiple source sentences into a closed semantic region by developing a *semantic constrained network* (SCN) in Section 3.1. And then introduce the proposed *controllable sampling* strategy in Section 3.2 to construct adequate and reasonable variations for source inputs.

3.1 Semantic Constrained Network

Network Architecture. One core component of our approach is the proposed SCN, which aims to

learn the global semantics and make them no difference between multiple source sentences ($\mathbf{x}$ and $\bar{\mathbf{x}}$). We adopt the CNN to address the variable-length problem of a sequence of hidden representations $\mathbf{H_x}$ (which is the output of the top encoder layer given $\mathbf{x}$) of the encoder stack. Formally, given an encoded representation $\mathbf{H_x} = \mathbf{H}_{x_1}, \mathbf{H}_{x_2}, ..., \mathbf{H}_{x_{T_{\mathbf{x}}}}$, the SCN first represents it as:

$$\xi_{1:T_{\mathbf{x}}} = \mathbf{H}_{x_1} \oplus \mathbf{H}_{x_2} \oplus ... \oplus \mathbf{H}_{x_{T_{\mathbf{x}}}}, \quad (3)$$

where $\oplus$ is the concatenation operator to build the matrix $\xi_{1:T_{\mathbf{x}}}$. Then a convolution operation involves a kernel $\mathbf{W}_c$ is applied to a window of l words to produce a new feature:

$$c_i = \text{Relu}(\mathbf{W}_c \otimes \xi_{i:i+l-1} + \mathbf{b}), \quad (4)$$

where $\otimes$ operator is the summation of element-wise production, $\mathbf{b}$ is a bias term. Finally we apply a max-over-time pooling operation over the feature map $\mathbf{c} = \max\{c_1, c_2, .., c_{T_{\mathbf{x}}-l+1}\}$ to capture the most important feature, that is, one with the highest value. We can use various numbers of kernels with different window sizes to repeat the above process, and extract different features to form the semantic representation, denoted as $\mathbf{H_c}$ for $\mathbf{x}$ (and $\mathbf{H_{\bar{c}}}$ for $\bar{\mathbf{x}}$ in a symmetric way).

Semantic Agreement Training. Given the semantic distributions $\mathcal{P}_\phi(\mathbf{z}|\mathbf{x})$ of $\mathbf{x}$ and $\mathcal{P}_\phi(\bar{\mathbf{z}}|\bar{\mathbf{x}})$ of $\bar{\mathbf{x}}$, we formulate $\ell_{\text{sem}}(\bar{\mathbf{x}}, \mathbf{x})$ as the negative **KL** divergence between them:

$$\ell_{\text{sem}}(\bar{\mathbf{x}}, \mathbf{x}) = -\mathbf{KL}\big(\mathcal{P}_\phi(\bar{\mathbf{z}}|\bar{\mathbf{x}})||\mathcal{P}_\phi(\mathbf{z}|\mathbf{x})\big). \quad (5)$$

We assume $\mathcal{P}_\phi(\mathbf{z}|\mathbf{x})$ and $\mathcal{P}_\phi(\bar{\mathbf{z}}|\bar{\mathbf{x}})$ have the following forms:

$$\mathcal{P}_\phi(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mu, \sigma^2\mathbf{I}), \; \mathcal{P}_\phi(\bar{\mathbf{z}}|\bar{\mathbf{x}}) \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2\mathbf{I}). \quad (6)$$

The mean μ ($\bar{\mu}$) and s.d. σ ($\bar{\sigma}$) are the outputs of neural networks based on the observation $\mathbf{H_c}$ (or $\mathbf{H_{\bar{c}}}$), as

$$\begin{aligned} \mu &= \mathbf{H_c} \cdot \mathbf{W}_\mu + \mathbf{b}_\mu, \\ \log \sigma^2 &= \mathbf{H_c} \cdot \mathbf{W}_\sigma + \mathbf{b}_\sigma, \end{aligned} \quad (7)$$

or

$$\begin{aligned} \bar{\mu} &= \mathbf{H_{\bar{c}}} \cdot \mathbf{W}_\mu + \mathbf{b}_\mu, \\ \log \bar{\sigma}^2 &= \mathbf{H_{\bar{c}}} \cdot \mathbf{W}_\sigma + \mathbf{b}_\sigma, \end{aligned} \quad (8)$$

where $\mathbf{W}_\mu, \mathbf{b}_\mu, \mathbf{W}_\sigma$ and $\mathbf{b}_\sigma$ are trainable parameters. To obtain a representation for latent semantic

Threshold	Method
$\hbar = 0$	Multinomial sampling
$\hbar = +\infty$	Greedy search
$\hbar \in (0, +\infty)$	Controllable sampling

Table 1: Multinomial sampling and greedy search, as special cases covered in controllable sampling.

distributions, we employ reparameterization technique as in (Kingma et al., 2014; Zhang et al., 2016). Formally,

$$\mathbf{z} = \mu + \sigma \odot \epsilon, \; \bar{\mathbf{z}} = \bar{\mu} + \bar{\sigma} \odot \epsilon, \quad (9)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ plays a role of introducing noises, and $\odot$ denotes an element-wise product. There can be other proper strategies to unify semantics of diverse inputs, we just present one example. Actually, the Gaussian form adopted here has several advantages, such as analytical evaluation of the KL divergence and ease of reparametrization for efficient gradient computation.

Augment Semantically. Given the encoder output $\mathbf{H_x}$ of $\mathbf{x}$, we augment it semantically with the captured semantics $\mathbf{z}$ by combining them with a gate $g = \text{sigmoid}(\mathbf{z} \cdot \mathbf{W}_{gz} + \mathbf{H}_{x_t} \cdot \mathbf{W}_{gx})$,

$$\mathbf{H_{o}}_t = \text{LayerNorm}(g \cdot \mathbf{z} + (1 - g) \cdot \mathbf{H}_{x_t}). \quad (10)$$

Identically, $\mathbf{H_{\bar{o}}}$ can be formulated given $\bar{\mathbf{z}}$ and $\mathbf{H_{\bar{x}}}$. Finally, the augmented source representation $\mathbf{H_o}$ (or $\mathbf{H_{\bar{o}}}$) is fed to the decoder to generate the final translation $\mathbf{y}$ conditioned on $\mathbf{x}$ (or $\bar{\mathbf{x}}$). In this strategy, our model can intuitively work well when meeting infrequent literal expressions as that can be pivoted by their corresponding semantic regions.

3.2 Controllable Sampling

For each target sentence $\mathbf{y}$, we need a set of reasonable source sentences $\mathcal{X}(\mathbf{y})$ to play the role of the inherent uncertainty. Unfortunately, it is extremely cost to annotate multiple source sentences manually for tens of million target sentences. To this end, we automatically construct $\mathcal{X}(\mathbf{y})$ using a well-trained *target-to-source* model $\overleftarrow{\theta}$ by sampling from the predicted word distributions:

$$\bar{\mathbf{x}}_t \sim P(\cdot|\bar{\mathbf{x}}_{<t}, \mathbf{y}; \overleftarrow{\theta}). \quad (11)$$

However, it is problematic to force the generation of a certain number of source sentences indiscriminately for each target sentence using beam

2727

search or multinomial sampling. The reason is that both of them synthesize sentences are either of less diverse or of less quality. Therefore, we propose a controllable sampling strategy to generate reasonable source sentences: at each decoding step, if the word distribution is sharp then we take the word with the maximum probability, otherwise the sampling method formulated in Eq. (11) is applied. Formally,

$$
\begin{cases}
\bar{\mathbf{x}}_t \sim P(\cdot|\bar{\mathbf{x}}_{<t}, \mathbf{y}; \overleftarrow{\boldsymbol{\theta}}), \text{if } \varepsilon \geq \hbar \\
\bar{\mathbf{x}}_t = \operatorname{argmax}\big(P(\cdot|\bar{\mathbf{x}}_{<t}, \mathbf{y}; \overleftarrow{\boldsymbol{\theta}})\big), \text{else}
\end{cases} \quad (12)
$$

where ε is exactly the information entropy respect to $P(\cdot|\bar{\mathbf{x}}_{<t}, \mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$:

$$
\begin{aligned}
\varepsilon = -\sum\nolimits_j \big[P(x^j|\bar{\mathbf{x}}_{<t}, \mathbf{y}; \overleftarrow{\boldsymbol{\theta}}) \times \\
\log P(x^j|\bar{\mathbf{x}}_{<t}, \mathbf{y}; \overleftarrow{\boldsymbol{\theta}}) \big],
\end{aligned} \quad (13)
$$

where $P(x^j|\bar{\mathbf{x}}_{<t}, \mathbf{y}; \overleftarrow{\boldsymbol{\theta}})$ denotes the conditional probability of the j-th word in the vocabulary appearing after the sequence $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{t-1}$. Actually, the widely used multinomial sampling and greedy search strategies can be served as special cases of the controllable sampling. $\hbar$ is a hyperparameter that indicates the sharpness threshold of the predicted word distributions and relates our method with the special cases as shown in Table 1. In practice, we repeat the above process N times to generate multiple source sentences to form $\mathcal{X}(\mathbf{y})$.

3.3 Training

Our framework initializes the model based on the parameters trained by the standard maximum likelihood estimation (MLE) (Eq. (1)). As shown in Eq. (2), the training objective of our approach is differentiable, which can be optimized using standard mini-batch stochastic gradient ascent techniques. To avoid the KL collapse (Bowman et al., 2016; Zhao et al., 2017), we use a simple scheduling strategy that sets $\gamma = 0$ at the beginning of training and gradually increases γ until $\gamma = 1$ is reached.

4 Experiments

We examine our method upon advanced TRANSFORMER (Vaswani et al., 2017) and conduct experiments on four widely used translation tasks, including WMT14 English→French (En→Fr), WMT16 English→German (En→De), NIST Chinese→English (Zh→En) and WMT18 Chinese→English.

4.1 Experimental Setting

Dataset For En→De, we used the WMT16[2] corpus containing 4.5M sentence pairs with 118M English words and 111M German words. The validation set is the concatenation of newstest2012 and newstest2013, and the results are reported on newstest2014 (test14), newstest2015 (test15) as well as newstest2016 (test16). For En→Fr, we used the significantly larger WMT 2014 English-French dataset consisting of 36M sentences. The validation set is the concatenation of newstest2012 and newstest2013, and the results are reported on newstest2014 (test14). For NIST Zh→En, we used the LDC[3] corpus consisting of 1.25M sentence pairs with 27.9M Chinese words and 34.5M English words respectively. We selected the best model using the NIST 2002 as the validation set for model selection and hyperparameters tuning. The NIST 2004 (MT04), 2005 (MT05), 2006 (MT06), and 2008 (MT08) datasets are used as test sets. For WMT18 Zh→En, we used a subset of WMT18 corpus containing 8M sentence pairs. We used newsdev 2017 as the validation set and reported results on newstest 2017 as well as newtest 2018.

We used the Stanford segmenter (Tseng et al., 2005) for Chinese word segmentation and applied the script `tokenizer.pl` of Moses (Koehn et al., 2007) for English, French and German tokenization. For En→De and En→Fr, all data had been jointly byte pair encoded (BPE) (Sennrich et al., 2016b) with 32k merge operations, which results in a shared source-target vocabulary. For NIST Zh→En, we created shared BPE codes with 60K operations that induce two vocabularies with 47K Chinese sub-words and 30K English sub-words. For WMT18 Zh→En, we used byte-pair-encoding to preprocess the source and target sentences, forming source- and target-side dictionaries with 32K types, respectively.

Model We adopt the `transformer_base` setting for Zh→En translations, while both `base` and `big` settings are adopted in En→De and En→Fr translations. For SCN, the filter windows are set to 2, 3, 4, 5 with 128 feature maps each. We set $\hbar = 2.5$, $N = 3$ for balancing the translation performance and the computation complexity. During training, we set $\lambda_1 = \lambda_2 = 0.5$, roughly 4,096 source and target to-

---

Method	Param.	Training Time (hours)	NIST Zh→En				WMT18 Zh→En	
			MT04	MT05	MT06	MT08	test17	test18
[†]Vaswani et al. (2017)	84M	9	47.37	46.81	46.34	38.23	24.41	24.59
Cheng et al. (2019)	N/A	N/A	49.13	49.04	47.74	38.61	N/A	N/A
TRANSFORMER	84M	9	47.14	47.03	46.26	38.31	24.69	24.61
TRANSFORMER$_\text{syn}$	84M	[‡]10	47.84	47.90	47.38	39.64	25.47	25.06
Ours	86M	[§]11.5	**49.15**	**49.21**	**48.88**	**40.94**	**26.48**	**26.36**

Table 2: BLEU [%] on Zh→En tasks. [†] denotes replicated results using tensor2tensor (T2T) toolkit. Both the training time and the number of parameters are related to the NIST Zh→En task. [‡]The time spent in synthesizing pseudo data was included. [§]Both the time spent in generating synthetic data and training models were included.

kens are paired in one mini-batch. We employ the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-9}$. Additionally, the same warmup and decay strategy for learning rate as Vaswani et al. (2017) is also used, with 8,000 warmup steps. For evaluation, we use beam search with a beam size of 4/5 and length penalty of 0.6/1.0 for En→{De,Fr}/Zh→En tasks respectively. We measure case-sensitive/insensitive tokenized BLEU[4] by `multi-bleu.pl`/`mteval-v11b.pl` for En→De/NIST Zh→En, while case-sensitive detokenized BLEU is reported by the official evaluation script `mteval-v13a.pl` for WMT18 Zh→En. Unless noted otherwise we run each experiment on up to four Tesla M40 GPUs and accumulate the gradients for 4 updates. For En→De/NIST Zh→En, each model was repeatedly run 4 times and we reported the average BLEU, while each model was trained only once on the larger WMT18 Zh→En dataset. For a strictly consistent comparison, we involve two strong baselines:

- **TRANSFORMER**, which is trained on the real parallel data only.

- **TRANSFORMER$_\text{syn}$**, which *is trained on the same data as ours* that consists of the real parallel data and the *back-translated* corpora. The latter contains N semantically-equivalent source sentences for each target sentence. These synthetic corpora are generated by a well-trained reverse NMT model using the proposed *controllable sampling* (see 3.2).

4.2 Comparison and Results

Table 2 shows the results on Zh→En tasks. For NIST Zh→En, we first compare our approach with the TRANSFORMER model on which our model is built. As we can see, our method can bring substantial improvements, which achieves notable gain of

Method	En→De			En→Fr
	test14	test15	test16	test14
Transformer base model				
Vaswani et al. (2017)	27.30	N/A	N/A	38.10
[‡]Cheng et al. (2018)	28.09	32.47	36.75	40.21
Cheng et al. (2019)	28.34	N/A	N/A	N/A
TRANSFORMER	27.67	32.04	36.18	39.86
TRANSFORMER$_\text{syn}$	27.81	32.33	36.62	40.07
Ours	**28.57**	**32.95**	**37.11**	**41.27**
Transformer big model				
Gao et al. (2019)	29.70	N/A	N/A	N/A
Cheng et al. (2019)	30.01	N/A	N/A	N/A
Ours	**30.29**	**34.21**	**38.28**	**42.92**

Table 3: BLEU [%] on En→De and En→Fr translation tasks. [‡]denotes our replicated results.

+2.36 BLEU points on average. In addition, our best model also achieves superior results across test sets to existing systems. For a more challenging task, we also report the results on WMT18 Zh→En task in Table 2. Compared with strong baseline systems, we observe that our method consistently improves translation performance on both newstest2017 and newstest2018. These results indicate that the effectiveness of our approach cannot be affected by the size of datasets.

Table 3 shows the results on WMT16 En→De and WMT14 En→Fr translations. For En→De, when investigating semantic augmentation into NMT models, significant improvements over two baselines (up to +0.91 and +0.62 BLEU points on average respectively) can be observed. We also take existing NMT systems as comparison which use almost the same English-German corpus. Our best system outperforms the standard Transformer (Vaswani et al., 2017) with +1.27 BLEU on newstest2014. It worth mentioning that our method outperforms the advanced robust NMT systems (Cheng et al., 2018, 2019), which aim to construct anti-noise NMT models, with at least +0.23 BLEU and up to +0.48 BLEU improvements. On En→Fr, our method outperforms both the previous

N	NIST Zh→En	WMT18 Zh→En	WMT16 En→De
1	43.09	23.01	24.50
3	44.01	23.80	25.22
5	43.92	23.89	25.30
7	44.09	23.77	25.31
9	44.14	23.68	25.37

Table 4: Effect of various numbers of synthetic source sentences on validation sets.

models and the in-house baselines. To further verify our approach, we study it with respect to `big` models and compare it with two related methods (Cheng et al., 2019; Gao et al., 2019). We can observe that the proposed approach achieves the best results among all methods for the same number of hidden units.

4.3 Analysis

Effect of N. To determine the number of synthetic source sentences N in our system beforehand, we conduct experiments on Zh→En and En→De translation tasks to test how it affects the translation performance. We vary the value of N from 1 to 9 with 2 as step size and the results are reported on validation sets (Table 4). We can find that the translation performance achieves substantial improvement with N increasing from 1 to 3. However, with N set larger than 3, we get little improvement. To make a trade-off between the translation performance and the computation complexity, we set N as 3 in our experiments.

Effect of $\hbar$. The introduction of the hyperparameter $\hbar$ aims at acquiring the proper quantity of synthetic data. To investigate the effect of it, we quantify: (1) the diversity using the edit distance among the synthetic source sentences and (2) the quality using BLEU scores of synthetic source sentences, with respect to various values of $\hbar$.

For each target sentence in validation sets, we generate $N = 3$ synthetic source sentences using *controllable sampling*. Table 5 shows the results. The BLEU scores were computed regarding the multiple synthetic sentences as a document. As in (Imamura et al., 2018), the edit distances are computed for two cases: (1) SYN vs. REAL, the average distance between a synthetic source sentence (SYN) and the real source sentence (REAL). (2) SYN vs. SYN, the average distance among synthetic source sentences of a target sentence ($C_3^2 = 3$ combinations per target sentence). We can find that

$\hbar$	BLEU	Edit Distance SYN vs. REAL	SYN vs. SYN
NIST Zh→En			
BS-3	20.87	8.70	5.14
0.0	10.71	17.26	19.18
1.0	11.99	17.17	18.91
2.5	_17.60_	_12.80_	_12.38_
4.5	19.47	9.93	6.24
7.0	20.30	9.07	4.35
WMT18 Zh→En			
BS-3	34.47	10.74	4.73
0.0	24.01	22.55	21.41
1.0	25.22	22.03	21.09
2.5	_31.29_	_12.58_	_12.24_
4.5	32.96	9.31	6.29
7.0	33.81	9.37	5.24
WMT16 En→De			
BS-3	30.11	9.59	4.36
0.0	19.84	15.60	15.75
1.0	20.57	15.22	15.26
2.5	_26.44_	_10.45_	_10.23_
4.5	28.07	8.38	3.95
7.0	29.25	7.29	2.71

Table 5: Effect of $\hbar$ on validation sets with respect to BLEU scores as well as edit distances among synthetic and real source sentences. "*BS-3*" indicates that synthetic sentences are generated by beam search with a beam size of 3.

$\ell_{\mathrm{mle}}(\mathbf{x}, \mathbf{y})$	ℓ_{sem}	$\ell_{\mathrm{mle}}(\bar{\mathbf{x}}, \mathbf{y})$	BLEU
✓			22.59
✓		✓	23.37
✓	✓		23.68
✓	✓	✓	24.10

Table 6: Ablation study on WMT18 Zh→En validation set. "✓" means the loss function is included in the training objective.

when $\hbar$ tends to 0 our controlled sampling method achieves lowest BLEU scores but highest edit distances. However, if we increase $\hbar$ gradually, it can be quickly simplified to greedy search. Among all values of $\hbar$ in Table 5, $\hbar = 2.5$ is a proper setting as it demonstrates relatively higher BLEU scores and lower word error rates (SYN vs. REAL) as well as more of diversity (SYN vs. SYN) in corpora. Therefore, we set $\hbar$ as 2.5 in all of our experiments. In addition, we can observe that the controllable sampling achieves the goal of generating sentences are of both high diverse and high quality.

Ablation Study. We perform an ablation study of our training objective formulated in Eq. (2) that contains three loss items. As shown in Table 6, the translation performance decreases by

Ours Method	N	BLEU
w/ beam search	3	23.41
w/ Multinomial sampling	3	23.74
w/ Controllable sampling	3	24.10

Table 7: Effect of different methods to generate multiple synthetic data. Experiments are conducted on WMT18 Zh→En validation set.
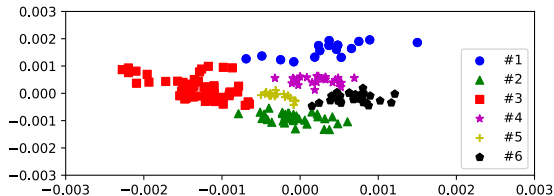

Figure 2: The figure shows a 2-dimensional PCA projection of the semantic representations ($\mathbf{z}$ in Eq. (9)) for six examples in the training corpus.

0.42 BLEU points when removing $\ell_{\mathrm{mle}}(\bar{\mathbf{x}}, \mathbf{y})$ while that increases to 0.73 BLEU points when ℓ_{sem} is excluded. In addition, only adding ℓ_{sem} is able to achieve an improvement of +1.09 BLEU points.

Effect of Controllable Sampling The widely used multinomial sampling and beam (greedy) search can be viewed as two special cases of the newly introduced controllable sampling. As in Table 7, our controllable sampling method achieves the best result among them on the validation set. We think that reasonable uncertainties can be mined via our controllable sampling strategy.

Visualization of Latent Space. We would like to verify whether our approach can capture semantics. Fortunately, there are such cases in the training set: a target sentence appears several times with different source sentences. We take some of them as examples, in which there are at least 17 unique source sentences for each target sentence. We visualize the semantic representations captured by the SCN of these examples in Figure 2. We observe that the representations are clearly clustered into 6 groups as expected, although demonstrating some noises, which reveal the strong capability of our approach to capture semantic representations.

Case Study. Table 8 shows an example translation (more examples are shown in the Appendix C). In this example, input #1, #2 and #3 have the same meaning. For input #1 and #3, "启动" and "上新" both mean "launch new products" here. The input #2 presents a different literal expression. Compared

Input #1	我认为我们可以重新启动这些品牌，而且现在时间正合适。
TRANS$_{\mathrm{syn}}$	I think we can restart these brands, and the time is right.
Ours	I think we can relaunch these brands, and now is the right time.
Input #2	我想现在是时候重新发布这些品牌了。
TRANS$_{\mathrm{syn}}$	I think it is time to reissue these brands.
Ours	I think it's time to relaunch these brands.
Input #3	我认为我们可以重新上新这些品牌，而且现在时间正合适。
TRANS$_{\mathrm{syn}}$	I think we can renew these brands, and now is the right time.
Ours	I think we can re-launch these brands, and the time is right now.

Table 8: Translation examples of TRANSFORMER$_{\mathrm{syn}}$ (TRANS$_{\mathrm{syn}}$ for short) and our method on various inputs under the same meaning on WMT18 Zh→En.

to TRANSFORMER$_{\mathrm{syn}}$, our approach motivates the models to be consistent when processing equivalent source inputs with various lteral expressions.

4.4 Semi-supervised Setting

Given the effectiveness of leveraging monolingual data in improving translation quality (Sennrich et al., 2016a), we further propose to improve our proposed model using target monolingual data on WMT16 En→De translation. Specifically, we augment the original parallel data of WMT16 corpus containing 4.5M sentence pairs by 24M[5] unique sentences randomly extracted from German monolingual newscrawl data. All of them are no longer than 100 words after tokenizing and BPE processing. We synthesize multiple source sentences for each monolingual sentence via controllable sampling (Section 3.2), and the one with the highest probability is served as the *real* source sentence (i.e., $\mathbf{x}$). We upsample the parallel data with a rate of 5 so that we observe every bitext sentence 5 times more often than each monolingual sentence. The resulted data is finally used to re-train our models and perform 300K updates on 8 P100 GPUs. Due to resource constraints, we adopt the smaller `transformer_base` setting here.

Table 9 summarizes our results and compares to other work in the literature. After incorporating monolingual data, our method yields an additional gain of +1.5~+3.3 BLEU points. For comparison,

Method	test14	test15	test16
Wang et al. (2019), `big`	31.00	32.01	N/A
Edunov et al. (2018), `big`	**35.00**	34.87	37.89
TRANSFORMER, `base`	27.67	32.04	36.18
+ Monolingual Data	30.14	34.17	37.28
Ours, `base`	28.57	32.95	37.11
+ Monolingual Data	31.87	**35.19**	**38.65**

Table 9: BLEU scores [%] on WMT16 En→De test sets (newstest2014∼2016) with monolingual data. Wang et al. (2019) used 2M extra back-translated data and Edunov et al. (2018) used 226M German monolingual sentences during back-translation.

Wang et al. (2019) quantify the prediction confidence using model uncertainty to alleviate the noisy back-translated parallel data and achieve 31 BLEU on newstest2014. Edunov et al. (2018) achieve as high as 35.0 BLEU on newstest2014 by adopting the `transformer_big` setting and relying on massive (scale to 226M) monolingual data. For comparison, our models fall behind Edunov et al. (2018)'s method on newstest2014 but achieve superior results on other two test sets. This reveals that the proposed method is surprisingly effective and complements existing non-semantic data augmentation techniques quite well.

5   Conclusion and Future Work

We present an uncertainty-aware semantic augmentation method to bridge the discrepancy of the data distribution between the training and the inference phases for dominant NMT models. In particular, we first synthesize a proper number of source sentences to play the role of intrinsic uncertainties via the controllable sampling for each target sentence. Then, we develop a semantic constrained network to summarize multiple source inputs into a closed semantic region which is then utilized to augment latent representations. Experiments on WMT14 English→French, WMT16 English→German, NIST Chinese→English and WMT18 Chinese→English translation tasks show that the proposed method can achieve consistent improvements across different language pairs.

While we showed that uncertainty-aware semantic augmentation with Gaussian priors is effective, more work is required to investigate if such an approach will also be successful for more sophisticated priors. In addition, learning universal representations among semantically-equivalent source and target sentences (Wei et al., 2020) can complete

the proposed method.

Acknowledgments

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In *Proceedings of ACL-08: HLT*, pages 200–208, Columbus, Ohio. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

Sufeng Duan, Hai Zhao, Dongdong Zhang, and Rui Wang. 2020. Syntax-aware data augmentation for neural machine translation. In *arXiv:2004.14200*.

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, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5539–5544, Florence, Italy. Association for Computational Linguistics.

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: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828.

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, Melbourne, Australia. Association for Computational Linguistics.

Kenji Imamura, Atsushi Fujita, and Eiichiro Sumita. 2018. Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63, Melbourne, Australia. Association for Computational Linguistics.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages

1681–1691, Beijing, China. Association for Computational Linguistics.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. 2017. Bayesian segnet: Model uncertainty in deep convolutional encoder–decoder architectures for scene understanding. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5574–5584.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3581–3589.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Guanlin Li, Lemao Liu, Guoping Huang, Conghui Zhu, and Tiejun Zhao. 2019. Understanding data augmentation in neural machine translation: Two perspectives towards generalization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5693–5699, Hong Kong, China. 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, Melbourne, Australia. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Granger, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3953–3962.

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 (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. 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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Harshil Shah and David Barber. 2018. Generative neural machine translation. In *Advances in Neural Information Processing Systems*, pages 1346–1355.

Aili Shen, Daniel Beck, Bahar Salehi, Jianzhong Qi, and Timothy Baldwin. 2019. Modelling uncertainty in collaborative document quality assessment. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 191–201, Hong Kong, China. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 791–802, Hong Kong, China. Association for Computational Linguistics.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Xiangpeng Wei, Yue Hu, Luxi Xing, Yipeng Wang, and Li Gao. 2019. Translating with bilingual topic knowledge for neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 7257–7264. AAAI Press.

Xiangpeng Wei, Yue Hu Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. 2020. On learning universal representations across languages. *CoRR*, abs/2007.15960.

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, Brussels, Belgium. Association for Computational Linguistics.

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

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2018. Quantifying uncertainties in natural language processing tasks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7322–7329.

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. 2017. Data noising as smoothing in neural network language models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Mingming Yang, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Min Zhang, and Tiejun Zhao. 2019. Sentence-level agreement for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics, pages 3076–3082, Florence, Italy. Association for Computational Linguistics.

Poorya Zaremoodi and Gholamreza Haffari. 2018. Incorporating syntactic uncertainty in neural machine translation with a forest-to-sequence model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1421–1429, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

Xiang Zhang, Shizhu He, Kang Liu, and Jun Zhao. 2019a. AdaNSP: Uncertainty-driven adaptive decoding in neural semantic parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4265–4270, Florence, Italy. Association for Computational Linguistics.

Xuchao Zhang, Fanglan Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2019b. Mitigating uncertainty in document classification. 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 3126–3136, Minneapolis, Minnesota. 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, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 555–562.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.