Improving Back-Translation with Uncertainty-based Confidence Estimation

Shuo Wang[†], **Yang Liu**[†⋆*], **Chao Wang**[⊤], **Huanbo Luan**[†], **and Maosong Sun**[†]
[†]Institute for Artificial Intelligence
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology
[⋆]Beijing Advanced Innovation Center for Language Resources
[⊤]6ESTATES PTE LTD, Singapore

Abstract

While back-translation is simple and effective in exploiting abundant monolingual corpora to improve low-resource neural machine translation (NMT), the synthetic bilingual corpora generated by NMT models trained on limited authentic bilingual data are inevitably noisy. In this work, we propose to quantify the confidence of NMT model predictions based on model uncertainty. With word- and sentence-level confidence measures based on uncertainty, it is possible for back-translation to better cope with noise in synthetic bilingual corpora. Experiments on Chinese-English and English-German translation tasks show that uncertainty-based confidence estimation significantly improves the performance of back-translation. [1]

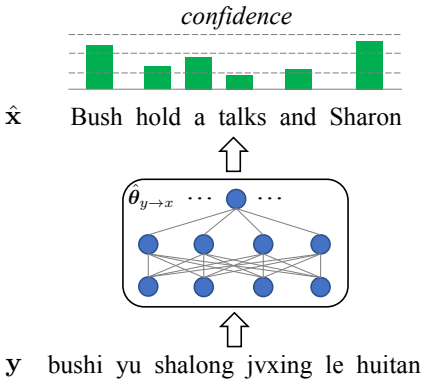

Figure 1: Confidence estimation for back-translation. Back-translation generates a source (e.g., English) sentence for a ground-truth target (e.g., Chinese) sentence. Such synthetic sentence pairs are used to train NMT models. As the model prediction (i.e., $\hat{\mathbf{x}}$) is often noisy, our work aims to quantify the prediction confidence using model uncertainty to alleviate error propagation.

1 Introduction

The past several years have witnessed the rapid development of end-to-end neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017), which leverages neural networks to map between natural languages. Capable of learning representations from data, NMT has significantly outperformed conventional statistical machine translation (SMT) (Koehn et al., 2003) and been widely deployed in large-scale MT systems in the industry (Wu et al., 2016; Hassan et al., 2018).

Despite the remarkable success, NMT suffers from the data scarcity problem. For most language pairs, large-scale, high-quality, and wide-coverage bilingual corpora do not exist. Even for the top handful of resource-rich languages, the major sources of available parallel corpora are often restricted to government documents or news articles.

Therefore, improving NMT under small-data training conditions has attracted extensive attention in recent years (Sennrich et al., 2016a; Cheng et al., 2016; Zoph et al., 2016; Chen et al., 2017; Fadaee et al., 2017; Ren et al., 2018; Lample et al., 2018). Among them, back-translation (Sennrich et al., 2016a) is an important direction. Its basic idea is to use an NMT model trained on limited authentic bilingual corpora to generate synthetic bilingual corpora using abundant monolingual data. The authentic and synthetic bilingual corpora are then combined to re-train NMT models. Due to its simplicity and effectiveness, back-translation has been widely used in low-resource language translation. However, as the synthetic corpora generated by the NMT model are inevitably noisy, translation errors can be propagated to subsequent steps and prone to hinder the

* Yang Liu is the corresponding author: liuyang2011@tsinghua.edu.cn.

[1]The source code is available at https://github.com/THUNLP-MT/UCE4BT

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 791–802, Hong Kong, China, November 3–7, 2019.

performance (Fadaee and Monz, 2018; Poncelas et al., 2018).

In this work, we propose a method to quantify the confidence of NMT model predictions to enable back-translation to better cope with translation errors. The central idea is to use *model uncertainty* (Buntine and Weigend, 1991; Gal and Ghahramani, 2016; Dong et al., 2018; Xiao and Wang, 2019) to measure whether the model parameters can best describe the data distribution. Based on the expectation and variance of word- and sentence-level translation probabilities calculated by Monte Carlo Dropout (Gal and Ghahramani, 2016), we introduce various confidence measures.

Different from most previous quality estimation studies that require feature extraction (Blatz et al., 2004; Specia et al., 2009; Salehi et al., 2014) or post-edited data (Kim et al., 2017; Wang et al., 2018; Ive et al., 2018) to train external confidence estimators, all our approach needs is the NMT model itself. Hence, it is easy to apply our approach to arbitrary NMT models trained for arbitrary language pairs. Experiments on Chinese-English and English-German translation tasks show that our approach significantly improves the performance of back-translation.

2 Background

Let $\mathbf{x} = x_1 \dots x_I$ be a source-language sentence and $\mathbf{y} = y_1 \dots y_J$ be a target-language sentence. We use $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_{x\to y})$ to denote a source-to-target NMT model (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) parameterized by $\boldsymbol{\theta}_{x\to y}$. Similarly, the target-to-source NMT model is denoted by $P(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}_{y\to x})$.

Let $\mathcal{D}_b = \{\langle \mathbf{x}^{(m)}, \mathbf{y}^{(m)} \rangle\}_{m=1}^{M}$ be an *authentic* bilingual corpus that contains M sentence pairs and $\mathcal{D}_m = \{\mathbf{y}^{(n)}\}_{n=1}^{N}$ be a monolingual corpus that contains N target sentences. The first step of back-translation (Sennrich et al., 2016a) is to train a target-to-source model on the authentic bilingual corpus $\mathcal{D}_b$ using maximum likelihood estimation:

$$\hat{\boldsymbol{\theta}}_{y\to x} = \underset{\boldsymbol{\theta}_{y\to x}}{\operatorname{argmax}} \Big\{ L(\mathcal{D}_b, \boldsymbol{\theta}_{y\to x}) \Big\}, \quad (1)$$

where the log-likelihood is defined as

$$L(\mathcal{D}_b, \boldsymbol{\theta}_{y\to x}) = \sum_{m=1}^{M} \log P(\mathbf{x}^{(m)}|\mathbf{y}^{(m)}, \boldsymbol{\theta}_{y\to x}). \quad (2)$$

The second step is to use the trained model $\hat{\boldsymbol{\theta}}_{y\to x}$ to translate the monolingual corpus $\mathcal{D}_m$:

$$\hat{\mathbf{x}}^{(n)} = \underset{\mathbf{x}}{\operatorname{argmax}} \Big\{ P(\mathbf{x}|\mathbf{y}^{(n)}, \hat{\boldsymbol{\theta}}_{y\to x}) \Big\}, \quad (3)$$

where $\hat{\mathbf{x}}^{(n)} = \hat{x}_1^{(n)} \dots \hat{x}_I^{(n)}$. The word-level decision rule is given by

$$\hat{x}_i^{(n)} = \underset{x}{\operatorname{argmax}} \Big\{ P(x|\mathbf{y}^{(n)}, \hat{\mathbf{x}}_{<i}^{(n)}, \hat{\boldsymbol{\theta}}_{y\to x}) \Big\}. \quad (4)$$

The resulting translations $\{\hat{\mathbf{x}}^{(n)}\}_{n=1}^{N}$ can be combined with $\mathcal{D}_m$ to generate a *synthetic* bilingual corpus $\tilde{\mathcal{D}}_b = \{\langle \hat{\mathbf{x}}^{(n)}, \mathbf{y}^{(n)} \rangle\}_{n=1}^{N}$.

The third step is to train a source-to-target model $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_{x\to y})$ on the combination of authentic and synthetic bilingual corpora:

$$\hat{\boldsymbol{\theta}}_{x\to y} = \underset{\boldsymbol{\theta}_{x\to y}}{\operatorname{argmax}} \Big\{ L(\mathcal{D}_b \cup \tilde{\mathcal{D}}_b, \boldsymbol{\theta}_{x\to y}) \Big\}. \quad (5)$$

This three-step process can iterate until convergence (Hoang et al., 2018; Cotterell and Kreutzer, 2018).

A problem with back-translation is that model predictions are inevitably erroneous. Translation errors can be propagated to subsequent steps and impair the performance of back-translation, especially when $\tilde{\mathcal{D}}_b$ is much larger than $\mathcal{D}_b$ (Pinnis et al., 2017; Fadaee and Monz, 2018; Poncelas et al., 2018). Therefore, it is crucial to develop principled solutions to enable back-translation to better deal with the error propagation problem.

3 Approach

This work aims to find solutions to the two following problems:

1. How to quantify the confidence of model predictions at both word and sentence levels?

2. How to leverage confidence to improve back-translation?

Section 3.1 introduces how to calculate model uncertainty, which lays a foundation for designing uncertainty-based word- and sentence-level confidence measures in Section 3.2. Section 3.3 describes confidence-aware training for NMT models on noisy bilingual corpora.

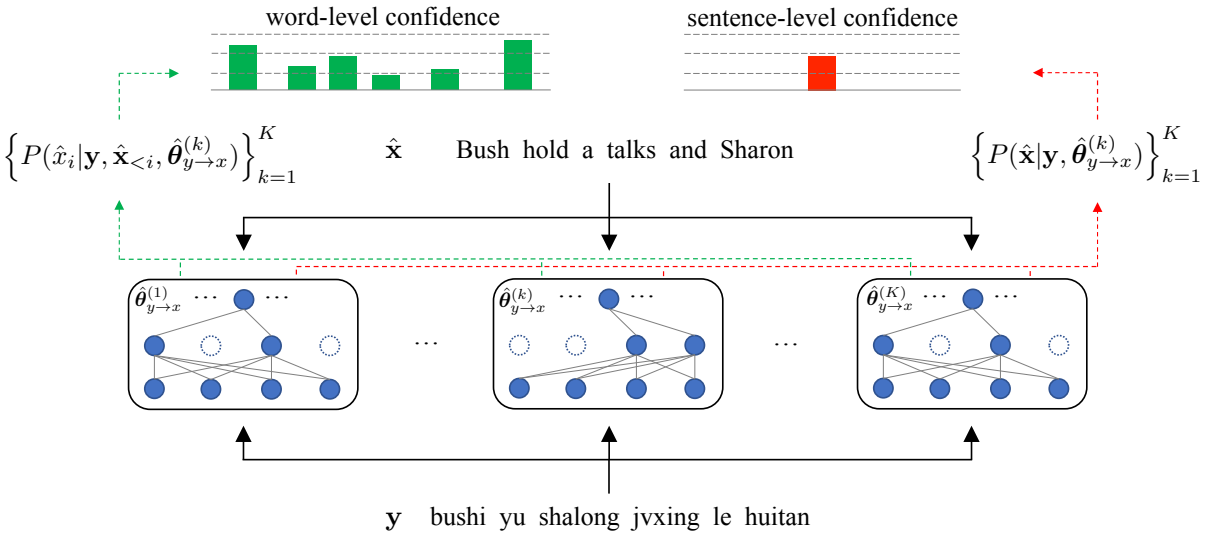

Figure 2: Illustration of uncertainty calculation. Given a target sentence $\mathbf{y}$ and the model prediction $\hat{\mathbf{x}}$, our approach treats word- and sentence-level translation probabilities as random variables and uses Monte Carlo Dropout to draw samples. These samples are used to calculate the expectations and variances of translation probabilities.

3.1 Calculating Uncertainty

Uncertainty quantification, which quantifies how confident a certain mapping is with respect to different inputs, has made significant progress due to the recent advances in Bayesian deep learning (Kendall et al., 2015; Gal and Ghahramani, 2016; Kendall and Gal, 2017; Xiao and Wang, 2019; Oh et al., 2019; Geifman et al., 2019; Lee et al., 2019). In this work, we aim to calculate *model uncertainty* (Kendall and Gal, 2017; Dong et al., 2018; Xiao and Wang, 2019), which measures whether a model can best describe the data distribution, for NMT using approximate inference methods widely used in Bayesian neural networks.

Given the authentic bilingual corpus $\mathcal{D}_b$, Bayesian neural networks aim at finding the posterior distribution over model parameters $P(\boldsymbol{\theta}_{y\to x}|\mathcal{D}_b)$. With a target sentence $\mathbf{y}$ in the monolingual corpus $\mathcal{D}_m$ and its translation $\hat{\mathbf{x}}$, the translation probability is given by

$$P(\hat{\mathbf{x}}|\mathbf{y}, \mathcal{D}_b) = \int P(\hat{\mathbf{x}}|\mathbf{y}, \boldsymbol{\theta}_{y\to x}) P(\boldsymbol{\theta}_{y\to x}|\mathcal{D}_b) d\boldsymbol{\theta}_{y\to x}. \quad (6)$$

In particular, we are interested in calculating the variance of the distribution $P(\hat{\mathbf{x}}|\mathbf{y}, \boldsymbol{\theta}_{y\to x})$ that reflects our ignorance over model parameters, which is referred to as *model uncertainty*. As exact inference is intractable, a number of variational inference methods (Graves, 2011; Blundell et al., 2015; Gal and Ghahramani, 2016) have been proposed to find an approximation to $P(\boldsymbol{\theta}_{y\to x}|\mathcal{D}_b)$. In this work, we leverage the widely used Monte Carlo Dropout (Gal and Ghahramani, 2016) to obtain samples of word- and sentence-level translation probabilities.

Figure 2 illustrates the key idea of our approach. Given an authentic target sentence $\mathbf{y}$, an NMT model made its prediction $\hat{\mathbf{x}}$ via a standard decoding process (see Eq. (3) and Eq. (4)). To quantify how confident the model was when making the prediction, our approach treats word- and sentence-level translation probabilities as random variables. [2] Drawing samples can be done by randomly deactivating part of neurons of the NMT model and re-calculating translation probabilities while keeping $\mathbf{y}$ and $\hat{\mathbf{x}}$ fixed. This stochastic feed-forward is repeated K times and generates K samples for both word- and sentence-level translation probabilities, respectively. We use $\hat{\boldsymbol{\theta}}^{(k)}_{y\to x}$ to denote the model parameters derived from $\hat{\boldsymbol{\theta}}_{y\to x}$ by deactivation in the k-th pass.

Intuitively, if the variance of translation probability is low, it is highly likely that the model was confident in making the prediction. Given K samples $\{P(\hat{\mathbf{x}}|\mathbf{y}, \hat{\boldsymbol{\theta}}^{(k)}_{y\to x})\}_{k=1}^{K}$, the expectation

[2] Unlike prior studies that calculate model uncertainty during inference (Xiao and Wang, 2019), our approach computes uncertainty after the NMT model has made the prediction for two reasons. First, our goal is to quantify the confidence of model prediction rather than using uncertainty to improve model prediction. Second, using Monte Carlo Dropout during decoding is very slow because of the autoregressive property of standard NMT models.

of sentence-level translation probability can be approximated by

$$\mathbb{E}\Big[P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x})\Big] \approx \frac{1}{K}\sum_{k=1}^{K} P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}^{(k)}_{y\to x}). \quad (7)$$

The variance of sentence-level translation probability can be approximated by

$$\begin{aligned}&\mathrm{Var}\Big[P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x})\Big]\\ &\approx \frac{1}{K}\sum_{k=1}^{K} P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}^{(k)}_{y\to x})^2 - \mathbb{E}\Big[P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x})\Big]^2, \quad (8)\end{aligned}$$

which is also referred to as *model uncertainty*.

The expectation and variance of word-level translation probabilities can also be calculated similarly using K samples.

3.2 Confidence Measures

We use $C(\mathbf{y},\hat{\mathbf{x}}_{<i},\hat{x}_i,\hat{\boldsymbol{\theta}}_{y\to x})$ to denote the *word-level confidence* for the model to generate $\hat{x}_i$ and $C(\mathbf{y},\hat{\mathbf{x}},\hat{\boldsymbol{\theta}}_{y\to x})$ to the denote the *sentence-level confidence* for the model to generate $\hat{\mathbf{x}}$.

Intuitively, when making predictions, the more confident an NMT model is, the higher expectation and lower variance of translation probability are. For comparison reasons, we used the following four types of confidence measures at the sentence level in our experiments:

1. *Predicted translation probability* (PTP). The translation probability of model prediction during standard decoding (Eq. (3)):

$$C_{\mathrm{PTP}}(\mathbf{y},\hat{\mathbf{x}},\hat{\boldsymbol{\theta}}_{y\to x}) = P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x}). \quad (9)$$

2. *Expected translation probability* (EXP). The expectation of translation probability:

$$C_{\mathrm{EXP}}(\mathbf{y},\hat{\mathbf{x}},\hat{\boldsymbol{\theta}}_{y\to x}) = \mathbb{E}\Big[P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x})\Big]. \quad (10)$$

3. *Variance of translation probability* (VAR). The variance of translation probability:

$$\begin{aligned}&C_{\mathrm{VAR}}(\mathbf{y},\hat{\mathbf{x}},\hat{\boldsymbol{\theta}}_{y\to x})\\ &= \Big(1 - \mathrm{Var}\Big[P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x})\Big]\Big)^{\alpha}. \quad (11)\end{aligned}$$

4. *Combination of expectation and variance* (CEV). The combination of expectation and variance:

$$\begin{aligned}&C_{\mathrm{CEV}}(\mathbf{y},\hat{\mathbf{x}},\hat{\boldsymbol{\theta}}_{y\to x})\\ &= \left(1 - \frac{\mathrm{Var}\big[P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x})\big]}{\mathbb{E}\big[P(\hat{\mathbf{x}}|\mathbf{y},\hat{\boldsymbol{\theta}}_{y\to x})\big]}\right)^{\beta}. \quad (12)\end{aligned}$$

where α and β are hyper-parameters to control the gap between confidence values of sentences of different quality. Larger values of α and β lead to bigger gaps. [3]

In Eq. (12), our approach tries to combine the merits of expectation and variance by using variance divided by expectation because smaller variance and bigger expectation are expected to result in higher confidence. There may exist more sophisticated ways to estimate prediction confidence using model uncertainty (Dong et al., 2018). As we find that the measures mentioned above are easy-to-implement and prove to be effective in our experiments, we leave the investigation of more complex confidence measures for future work.

The word-level confidence measures can be defined similarly.

3.3 Confidence-aware Training for NMT

We propose confidence-aware training for NMT to enable NMT to make better use of noisy data. Word- and sentence-level confidence measures are complementary: while word-level confidence can provide more fine-grained information than the sentence-level counterpart, it is unable to cope with word omission errors that can only be captured at the sentence level. As a result, our approach incorporates both word- and sentence-level confidence measures into the training process. [4]

Using Sentence-level Confidence

It is easy to integrate sentence-level confidence into back-translation by modifying the likelihood function in Eq. (5):

$$\begin{aligned}&L(\mathcal{D}_b \cup \tilde{\mathcal{D}}_b, \boldsymbol{\theta}_{x\to y})\\ =\ &\sum_{m=1}^{M} \log P(\mathbf{y}^{(m)}|\mathbf{x}^{(m)},\boldsymbol{\theta}_{x\to y}) +\\ &\sum_{n=1}^{N} C(\mathbf{y}^{(n)},\hat{\mathbf{x}}^{(n)},\hat{\boldsymbol{\theta}}_{y\to x}) \times\\ &\log P(\mathbf{y}^{(n)}|\hat{\mathbf{x}}^{(n)},\boldsymbol{\theta}_{x\to y}). \quad (13)\end{aligned}$$

[3] Note that all confidence measures are between 0 and 1. Clearly, both the expectation and variance of a probability are between 0 and 1. It can be proved that the variance of a probability is no greater than the corresponding expectation. As a result, $C_{\mathrm{CEV}}(\cdot)$ is also between 0 and 1.

[4] Instead of applying confidence estimation to the second pass of decoding (Luong et al., 2017), we directly integrate confidence scores into the training process. These two kinds of methods are complementary.

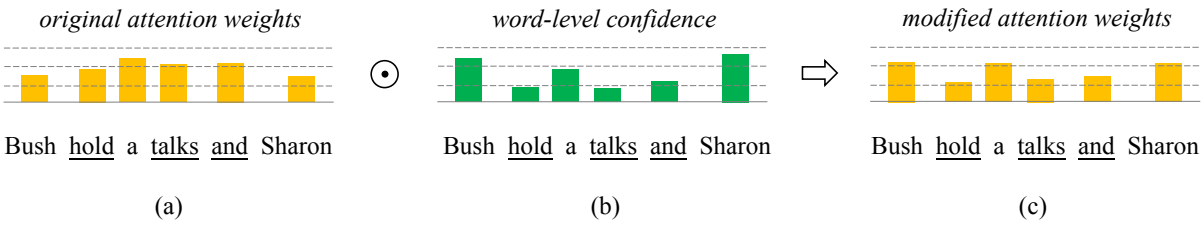

Figure 3: Using word-level confidence in confidence-aware training. The basic idea is to use confidence to modify attention weights to pay less attention to erroneous words highlighted in underline. (a) The original attention weights of the NMT model; (b) the word-level confidence of the noisy source sentence; (c) the attention weights modified by the word-level confidence, which focus more on words with high confidence. $\odot$ is a broadcast product. See Eq. (15) for details.

Serving as a weight assigned to each synthetic sentence pair, sentence-level confidence is expected to help to minimize the negative effect of estimating parameters on sentences with lower confidence. Note that the confidence of an authentic sentence pair in $\mathcal{D}_b$ is 1.

Using Word-level Confidence

As the source side instead of the target side of the synthetic bilingual corpus is noisy, word-level confidence cannot be integrated into back-translation in a similar way to sentence-level confidence. This is because the word-level confidence associated with each source word does not get involved in backpropagation during training.

Alternatively, we build a real-valued word-level confidence vector:

$$\mathbf{c} = \left\{ C\left(\mathbf{y}^{(n)}, \hat{\mathbf{x}}^{(n)}_{<i}, \hat{x}_i, \hat{\boldsymbol{\theta}}_{y \to x}\right) \right\}_{i=1}^{I}. \quad (14)$$

Due to the wide use of attention (Bahdanau et al., 2015; Vaswani et al., 2017) in NMT, we use the confidence vector $\mathbf{c} \in \mathbb{R}^{1 \times I}$ to modify attention weights and enable the model to focus more on words with high confidence. Figure 3 shows an example. Figure 3(a) gives a source sentence in the synthetic bilingual corpus, in which erroneous words "hold", "talks", and "and" receive high attention weights, deteriorating the parameter estimation on this sentence pair. By multiplying with word-level confidence (Figure 3(b)), the weights are modified to pay less attention to erroneous words (Figure 3(c)).

More formally, the modified attention function is given by

$$\begin{aligned} &\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{c}) \\ &= \left(\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}} \right) \odot \mathbf{c} \right) \mathbf{V}, \end{aligned} \quad (15)$$

where $\mathbf{Q} \in \mathbb{R}^{I \times D}$, $\mathbf{K} \in \mathbb{R}^{I \times D}$, and $\mathbf{V} \in \mathbb{R}^{I \times D}$ are query, key, and value matrices and D is the hidden size. $\odot$ is a broadcast product.

Since the integration of sentence- and word-level confidence measures are independent of each other, it is easy to use both of them in back-translation.

4 Experiments

4.1 Setup

We evaluated our approach on Chinese-English and English-German translation tasks. The evaluation metric is BLEU (Papineni et al., 2001) as calculated by the `multi-bleu.perl` script. We use the paired bootstrap resampling (Koehn, 2004) for significance testing.

For the Chinese-English task, the training set contains 1.25M sentence pairs from LDC[5] with 27.8M Chinese words and 34.5M English words. To build the monolingual corpus for back-translation, we extracted the English side of the training set of the WMT 2017 Chinese-English news translation task. After removing sentences longer than 256 words, we randomly selected 10M English sentences as the monolingual corpus. NIST06 is used as the development set and NIST02, 03, 04, 05, and 08 datasets as test sets.

For the English-German task, we used the dataset of the WMT 2014 English-German translation task. The training set consists of 4.47M sentence pairs with 116M English words and 110M German words. We randomly selected 4.5M German sentences from the 2012 News Crawl corpus of WMT 2014 to construct the monolingual corpus for back-translation. We use newstest 2013 as

[5]The training set includes LDC2002E18, LDC2003E07, LDC2003E14, part of LDC2004T07, LDC2004T08 and LDC2005T06.

Measure	BLEU	Δ
-	46.23	-
PTP	45.41	-0.82
EXP	45.22	-1.01
VAR	46.77	+0.54
CEV	**47.05**	**+0.82**

Table 1: Comparison of confidence measures.

Word	Sentence	BLEU	Δ
×	×	46.23	-
×	√	46.42	+0.19
√	×	46.98	+0.75
√	√	**47.05**	**+0.82**

Table 2: Comparison between word- and sentence-level CEV confidence measures.

the development set and newstest 2012, 2014, and 2015 as test sets.

Chinese sentences were segmented by an open-source toolkit THULAC[6] . German and English sentences were tokenized by the tokenizer in Moses (Koehn et al., 2007). We used byte pair encoding (Sennrich et al., 2016b) to perform sub-word segmentation with 32k merge operations for Chinese-English and 35k merge operations for English-German. Sentence pairs are batched together by approximate length and each batch has roughly 25,000 source and target tokens. We distinguish between three kinds of translations of the monolingual corpus:

1. NONE: there is no translation and only the authentic bilingual corpus is used;

2. SEARCH: the translations are generated by beam search (Sennrich et al., 2016a);

3. SAMPLE: the translations are generated by sampling (Edunov et al., 2018).

As neural quality estimation (Kim et al., 2017; Wang et al., 2018) can also give word- and sentence-level confidences for the output of NMT models when labeled data is available, we distinguish between two kinds of confidence estimation methods:

1. NEURALQE: the confidences are given by an external neural quality estimator;

2. UNCERTAINTY: the proposed uncertainty-based confidence estimation method.

For NEURALQE, we used the Predictor-Estimator architecture (Kim et al., 2017) implemented by OpenKiwi (Kepler et al., 2019), which is an open source software officially recommended by the QE shared task of WMT. Following the guide of OpenKiwi, we used a German-English parallel corpus containing 2.09M sentence pairs to train the predictor and a post-edited corpus containing 25k sentence triples to train the estimator. All the data used to train QE models are provided by WMT. As there are no post-edited corpora for the Chinese-English task, NEURALQE can only be used in the English-German task in our experiments. For NEURALQE, both word- and sentence- level quality scores were considered.

We implemented our method on the top of THUMT (Zhang et al., 2017). The NMT model we use is Transformer (Vaswani et al., 2017). We used the base model for the Chinese-English task and the big model for the English-German task. We used the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ to optimize model parameters. We used the same warm-up strategy for learning rate as Vaswani et al. (2017) with $\text{warmup\_steps} = 4,000$. During training, the hyper-parameter of label smoothing was set as $\epsilon_{ls} = 0.1$ (Szegedy et al., 2016; Pereyra et al., 2017). During training and the Monte Carlo Dropout process, the hyper-parameter of dropout was set to 0.1 and 0.3 for Transformer base and big models, respectively. K was set to 20. Through experiments, we find our method works best when the α and β are set to 2. All experiments were conducted on 8 NVIDIA GTX 1080Ti GPUs.

4.2 Comparison of Confidence Measures

Table 1 shows the comparison of confidence measures on the Chinese-English development set. We find that using either the translation probabilities outputted by the model (i.e., "PTP") or the expectation of translation probabilities (i.e., "EXP") deteriorates the translation quality, which suggests that translation probabilities themselves can not help NMT models better cope with synthetic data.

[6] https://github.com/thunlp/THULAC-Python

Data	CE	MT06	MT02	MT03	MT04	MT05	MT08	All
NONE	-	45.05	45.09	44.79	46.07	44.34	35.52	43.50
SEARCH	-	46.23	45.85	45.37	46.77	46.28	37.69	44.76
	U	**47.05**$^{++}$	**48.06**$^{++}$	**46.44**$^{++}$	**47.59**$^{++}$	**47.03**$^{++}$	**38.02**$^{+}$	**45.72**$^{++}$
SAMPLE	-	46.69	46.98	45.62	46.97	46.29	37.28	44.96
	U	**46.78**	46.75	**46.53**‡‡	**47.70**‡‡	**47.48**‡‡	36.99	**45.37**‡‡

Table 3: BLEU scores on the NIST Chinese-English translation task. The ratio of authentic data to synthetic data is 1:1. NONE: only the authentic bilingual corpus is used. SEARCH: the translations of the monolingual corpus are generated by beam search (Sennrich et al., 2016a). SAMPLE: the translations of the monolingual corpus are generated by sampling (Edunov et al., 2018). "CE": confidence estimation method. "U": the proposed uncertainty-based confidence estimation. "All": the combination of all test sets. "+": significantly better than SEARCH without CE ($p < 0.05$). "++": significantly better than SEARCH without CE ($p < 0.01$). "‡‡": significantly better than SAMPLE without CE ($p < 0.01$).

In contrast, using the variance or model uncertainty (i.e., "VAR") increases translation quality. Combining variance and expectation (i.e., "CEV") leads to a further improvement. In the following experiments, we use CEV as the default setting.

4.3 Comparison between Word- and Sentence-level Confidence Measures

Table 2 shows the comparison between word- and sentence-level CEV (i.e., combination of expectation and variance) confidence measures on the Chinese-English development set. It is clear that using either sentence-level or word-level confidence measures improves the translation performance. Thanks to more fine-grained quantification of uncertainty, using word-level confidence achieves a higher BLEU score than using sentence-level confidence. Combining the sentence- and word-level of confidences leads to a further improvement, suggesting that they are complementary to each other. In the following experiments, we use the combination of word- and sentence-level confidences as the default setting.

4.4 Main Results

The Chinese-English Task

Table 3 shows the results of the Chinese-English task. Back-translation, either generating translations using beam search (i.e., SEARCH) or using sampling (i.e., SAMPLE), does lead to significant improvements over using only the authentic bilingual corpus (i.e., NONE). We find that SAMPLE is more effective than SEARCH, which confirms the finding of Edunov et al. (2018). Using uncertainty-based confidence (i.e., "U") significantly improves over both SEARCH and SAMPLE on the combination of all test sets ($p < 0.01$). As there is no Chinese-English labeled data to train neural quality estimation models, we did not report the result of NEURALQE in this experiment.

The English-German Task

Table 4 shows the results of the English-German task. We find that using quality estimation, either NEURALQE (i.e., "N") or UNCERTAINTY (i.e., "U"), improves over SEARCH and SAMPLE. UNCERTAINTY even achieves better performance than NEURALQE, although NEURALQE uses additional labeled training data. As NEURALQE heavily relies on post-edited corpora and labeled data to train QE models, it can only be used in a handful of language pairs. In contrast, it is easier to apply our approach to arbitrary language pairs since it does not need any labeled data to estimate confidence.

4.5 Effect of Training Corpus Size

Figure 4 shows the effect of training corpus size. The X-axis is the size of the total training data (i.e., $\mathcal{D}_b \cup \tilde{\mathcal{D}}_b$ in Eq. (5)). The BLEU scores were calculated on the Chinese-English development set. We find that the translation performance of SEARCH rises with the increase of monolingual corpus size in the beginning. However, further enlarging the monolingual corpus hurts the translation performance. In contrast, our approach can still obtain further improvements when adding more synthetic bilingual sentence pairs. Similar findings are also observed for SAMPLE.

Data	CE	news2013	news2012	news2014	news2015	All
NONE	-	26.57	22.09	28.42	30.26	26.31
SEARCH	-	27.09	23.10	29.45	30.10	27.04
	N	27.58	23.91	30.61	31.87	28.18
	U	**27.89**$^{++*}$	23.75^{++}	**31.00**$^{++*}$	**31.98**$^{++}$	**28.28**$^{++}$
SAMPLE	-	27.30	23.37	30.11	31.51	27.70
	N	27.55	23.53	30.13	31.87	27.87
	U	**27.71**‡‡	**23.80**‡	**30.54**‡‡††	**32.01**‡	**28.15**‡‡††

Table 4: BLEU scores on the WMT14 English-German translation task. The ratio of authentic data to synthetic data is 1:1. NONE: only the authentic bilingual corpus is used. SEARCH: the translations of the monolingual corpus are generated by beam search (Sennrich et al., 2016a). SAMPLE: the translations of the monolingual corpus are generated by sampling (Edunov et al., 2018). "CE": confidence estimation method. "U": uncertainty-based confidence estimation. "N": NEURALQE. "All": the combination of all test sets. "++": significantly better than SEARCH without CE ($p < 0.01$). "$*$": significantly better than "SEARCH + N" ($p < 0.05$). "‡": significantly better than SAMPLE without CE ($p < 0.05$). "‡‡": significantly better than SAMPLE without CE ($p < 0.01$). "††": significantly better than "SAMPLE + N" ($p < 0.01$).

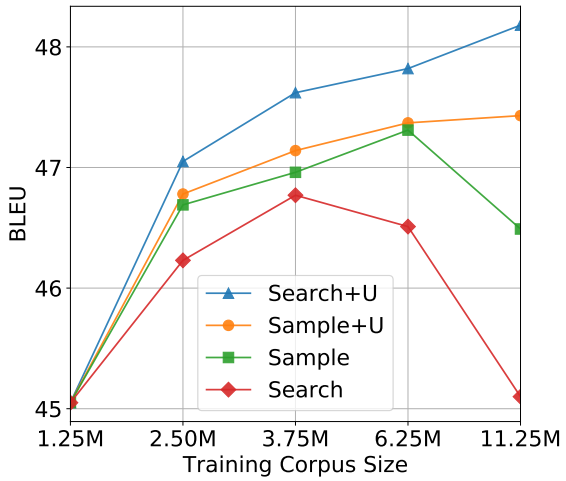

Figure 4: Effect of training corpus size.

4.6 Effect of Data Selection

Instead of randomly selecting monolingual sentences to generate synthetic data, we also used the method proposed by (Fadaee and Monz, 2018) to select monolingual data by targeting difficult words. In this series of experiments, we used the same amount of monolingual data that was derived from a larger monolingual corpus using different data selection methods.

Results on NIST06 show that targeting difficult words improves over randomly selecting monolingual data (46.23 → 46.60 BLEU), confirming the finding of Fadaee and Monz (2018). Using uncertainty-based confidence can further improve the translation performance (46.60 → 47.18 BLEU), indicating that our approach can be combined with advanced data selection methods.

4.7 Case Study

Figure 5 shows an example of model prediction and its corresponding word- and sentence-level confidence measures for the English-German task. We observe that the PTP and EXP measures are unable to give low confidence to erroneous words. In contrast, variance-based measures such as VAR and CEV can better quantify how confident the model is to make its prediction.

5 Related work

Our work is closely related to three lines of research: (1) back-translation, (2) confidence estimation, and (3) uncertainty quantification.

5.1 Back-translation

Back-translation is a simple and effective approach to leveraging monolingual data for NMT (Sennrich et al., 2016a). There has been a growing body of literature that analyzes and extends back-translation recently. Currey et al. (2017) show that low-resource NMT can benefit from the synthetic data generated by simply copying target monolingual data to the source side. Imamura et al. (2018) and Edunov et al. (2018) demonstrate that it is more effective to generate source sentences via sampling rather than beam search. Cotterell and Kreutzer (2018) and Hoang et al.

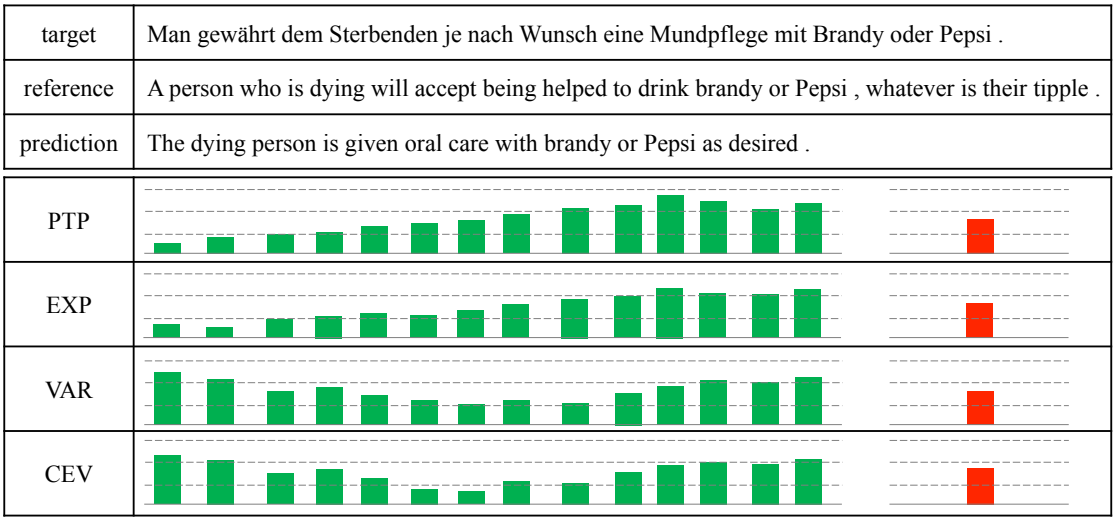

Figure 5: Example of confidence measures.

(2018) find that iterative back-translation can further improve the performance of NMT. Fadaee and Monz (2018) show that words with high predicted loss during training benefit most. Our work differs from existing methods in that we propose to use confidence estimation to enable back-translation to better cope with noisy synthetic data, which can be easily combined with previous works. Our experiments show that both neural and uncertainty-based confidence estimation methods benefit back-translation.

5.2 Confidence Estimation

Estimating the confidence or quality of the output of MT systems (Ueffing and Ney, 2007; Specia et al., 2009; Bach et al., 2011; Salehi et al., 2014; Rikters and Fishel, 2017; Kepler et al., 2019) is important for enabling downstream applications such as post-editing and interactive MT to better cope with translation mistakes. While existing methods rely on external models to estimate confidence, our approach leverages model uncertainty to derive confidence measures. The major benefit is that our approach does not need labeled data.

5.3 Uncertainty Quantification

Reliable uncertainty quantification is key to building a robust artificial intelligent system. It has been successfully applied to many fields, including computer vision (Kendall et al., 2015; Kendall and Gal, 2017), time series prediction (Zhu and Laptev, 2017), and natural language processing (Dong et al., 2018; Xiao and Wang, 2019). Our work differs from previous work in that we are interested in calculating uncertainty after the model has made the prediction rather during inference. Ott et al. (2018) also analyze the inherent uncertainty of machine translation. The difference is that they focus on the existence of multiple correct translations for a single sentence while we aim to quantify the uncertainty of NMT models.

6 Conclusions

We have presented a method for qualifying model uncertainty for neural machine translation and use uncertainty-based confidence measures to improve back-translation. The key idea is to use Monte Carlo Dropout to sample translation probabilities to calculate model uncertainty, without the need for manually labeled data. As our approach is transparent to model architectures, we plan to further verify the effectiveness of our approach on other downstream applications of NMT such as post-editing and interactive MT in the future.

Acknowledgments

We thank all anonymous reviewers for their valuable comments. This work is supported by the National Key R&D Program of China (No. 2017YFB0202204), National Natural Science Foundation of China (No. 61761166008, No. 61432013), Beijing Advanced Innovation Center for Language Resources (No. TYR17002), and the NExT++ project supported by the National Research Foundation, Prime Ministers Office, Singapore under its IRC@Singapore Funding Initiative. This research is also supported by Sogou Inc.

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 211–219, Portland, Oregon, USA. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. In *Proceedings of ICML 2015*.

Wray Buntine and Andreas S. Weigend. 1991. Bayesian back-propagation. *Complex Systems*.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. 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 ACL 2016*.

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

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156. Association for Computational Linguistics.

Li Dong, Chris Quirk, and Mirella Lapata. 2018. Confidence modeling for neural semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Melbourne, Australia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP 2018*.

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 EMNLP 2018*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approaximation: Representing model uncertainty in deep learning. In *Proceedings of ICML 2016*.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. 2019. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*.

Alex Graves. 2011. Practical variational inference for neural networks. In *Proceedings of NeurIPS*.

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. arXiv:1803.05567.

Cong Duy Vu Hoang, Philipp Keohn, 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*.

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. Association for Computational Linguistics.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. deepQuest: A framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv:1511.02680.

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*, pages 5574–5584.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 388–395.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL 2003*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of EMNLP 2018*.

Gilwoo Lee, Brian Hou, Aditya Mandalika, Jeongseok Lee, and Siddhartha S. Srinivasa. 2019. Bayesian policy optimization for model uncertainty. In *International Conference on Learning Representations*.

Ngoc-Quang Luong, Laurent Besacier, and Benjamin Lecouteux. 2017. Find the errors, get the better: Enhancing machine translation via word confidence estimation. *Natural Language Engineering*, 23:617–639.

Seong Joon Oh, Andrew C. Gallagher, Kevin P. Murphy, Florian Schroff, Jiyan Pan, and Joseph Roth. 2019. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of ICML 2018*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2001*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved subword units and synthetic data. In *Text, Speech, and Dialogue*, pages 237–245, Cham. Springer International Publishing.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *CoRR*, abs/1804.06189.

Shuo Ren, Wenhu Chen, Shujie Liu, Mu Li, Ming Zhou, and Shuai Ma. 2018. Triangular architecture for rare language translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 56–65, Melbourne, Australia. Association for Computational Linguistics.

Matiss Rikters and Mark Fishel. 2017. Confidence through attention. *CoRR*, abs/1710.03743.

Marzieh Salehi, Shahram Khadivi, and Nooshin Riahi. 2014. Confidence estimation for machine translation using context vectors. *7'th International Symposium on Telecommunications (IST'2014)*, pages 524–528.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of ACL 2016*.

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.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the confidence of machine translation quality estimates. In *Twelfth Machine Translation Summit*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NeurIPS 2014*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Nicola Ueffing and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*.

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 *Proceedings of NeurIPS 2017*.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for WMT18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 809–815, Belgium, Brussels. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144v2.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of AAAI 2019*.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huan-Bo Luan, and Yang Liu. 2017. THUMT: an open source toolkit for neural machine translation. *CoRR*, abs/1706.06415.

Lingxue Zhu and Nikolay Laptev. 2017. Deep and confident prediction for time series at uber. In *Proceedings of ICDM 2017*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP 2016*.