Reducing Length Bias in Scoring Neural Machine Translation via a Causal Inference Method

Xuewen Shi[1,2], **Heyan Huang**[1,2], **Ping Jian**[1,2]*, **Yi-Kun Tang**[1,2]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Beijing Engineering Research Center of
High Volume Language Information Processing and Cloud Computing Applications
{xwshi, hhy63, pjian, tangyk}@bit.edu.cn

Abstract

Neural machine translation (NMT) usually employs beam search to expand the searching space and obtain more translation candidates. However, the increase of the beam size often suffers from plenty of short translations, resulting in dramatical decrease in translation quality. In this paper, we handle the length bias problem through a perspective of causal inference. Specifically, we regard the model generated translation score S as a degraded true translation quality affected by some noise, and one of the confounders is the translation length. We apply a Half-Sibling Regression method to remove the length effect on S, and then we can obtain a debiased translation score without length information. The proposed method is model agnostic and unsupervised, which is adaptive to any NMT model and test dataset. We conduct the experiments on three translation tasks with different scales of datasets. Experimental results and further analyses show that our approaches gain comparable performance with the empirical baseline methods.

1 Introduction

Recently, with the renaissance of deep learning, end-to-end neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) has gained remarkable performances (Wu et al., 2016; Gehring et al., 2017; Vaswani et al., 2017). NMT models are usually built upon an encoder-decoder framework (Cho et al., 2014): the encoder reads an input sequence $\mathbf{x} = \{x_1, ..., x_{T_x}\}$ into a hidden memory H, and the decoder is designed to model a probability over the translation $\mathbf{y} = \{y_1, ..., y_{T_{\hat{\mathbf{y}}}}\}$ by:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T_y} P(y_t|y_{<t}, H). \tag{1}$$

Most existing NMT approaches employ beam search to obtain more translation candidates and then gain a better translation hypothesis $\hat{\mathbf{y}} = \{\hat{y}_1, \cdots, \hat{y}_{T_{\hat{\mathbf{y}}}}\}$ by ranking the translation candidates set $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_b\}$ across a score function $s(\hat{\mathbf{y}})$:

$$s(\hat{\mathbf{y}}) = \sum_{t=1}^{T_{\hat{\mathbf{y}}}} \log P(\hat{y}_t|\mathbf{x}; \theta), \tag{2}$$

where b is the beam size and θ is the parameter set of the NMT model.

However, continuously increasing the beam size has been shown to degrade performances and lead to short translations (Koehn and Knowles, 2017). One decisive reason is that the large search space is easy to introduce more short $\hat{\mathbf{y}}$, and the shorter $\hat{\mathbf{y}}$ tends to be scored higher under $s(\hat{\mathbf{y}})$ in Eq. (2). Previous efforts usually deal with the above length bias problem by two mechanisms: i) performing length normalization on $s(\hat{\mathbf{y}})$ via dividing $s(\hat{\mathbf{y}})$ by the length penalty lp, i.e. $s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}})/lp$ (Boulanger-Lewandowski et al., 2013; Jean et al., 2015; Koehn and Knowles, 2017; Yang et al., 2018; Meister et al., 2020), and ii) adding an additional length-related reward r to $s(\hat{\mathbf{y}})$, i.e. $s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) + \gamma \cdot r$ (Li and Jurafsky, 2016; He et al., 2016; Murray and Chiang, 2018; Huang et al., 2017; Yang et al., 2018). For

*Corresponding author: Ping Jian

Proceedings of the 20th China National Conference on Computational Linguistics, pages 874-885, Hohhot, China, Augest 13 - 15, 2021.

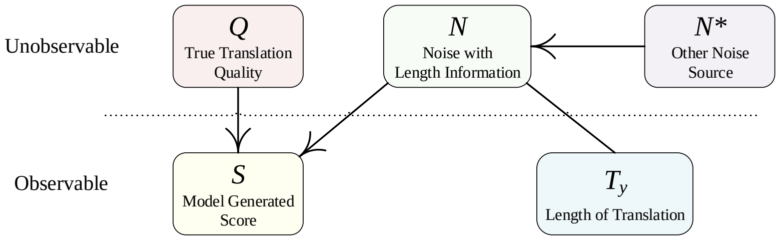


Figure 1: A causal directed acyclic graph shows the relations among the true translation quality Q, the model generated score S and the translation length $T_{\hat{\mathbf{y}}}$. See section 1 and section 3.1 for more details.

the second strategy, the correcting ratio γ of the reward is usually determined by supervised training (He et al., 2016; Murray and Chiang, 2018) or manually fine-tuning (Huang et al., 2017) before the testing stage, which lacks the ability of self-adapting to the unseen data.

In this paper, we introduce a causal motivated model agnostic and unsupervised method to solve the length bias problem for NMT. As shown in Fig. 1, for a translation hypothesis $\hat{\mathbf{y}}$, suppose that Q is an unobservable true translation quality of $\hat{\mathbf{y}}$, and the model generated score S can be seen as an observed degraded version of Q which is affected by some noise N. Generally, S equals $s(\hat{\mathbf{y}})$ in conventional NMT approaches, and it can be viewed as one of the measurement methods of Q with systematic errors. As mentioned above, one kind of systematic errors has a strong correlation with the translation length, therefore, the noise caused by length will be eliminated if we subtract the length effect from S. Specifically, we utilize the Half-Sibling Regression (HSR) (Schölkopf et al., 2016) method to perform the noise elimination operation for NMT. The method first apply a regression model to appraise the effect of the translation length on the model generated score, i.e. $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$. Then, the denoised score is obtained by removing $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ from S:

$$S' := S - \mathrm{E}[S|T_{\hat{\mathbf{y}}}]. \tag{3}$$

We propose two branches of the framework, corpus based (C-HSR) and single source sentence based (S-HSR) re-scoring method. The difference is that C-HSR performs the estimation of $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ on the whole test set, while S-HSR uses the translation candidates in a beam of the NMT inference process to predict $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$. The operation of approximating $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ for both C-HSR and S-HSR entirely rely on the current testing data of NMT without fine-tuning or any supervised information. In this work, we regard the NMT model as a black-box and apply the HSR-based denoised method to the re-scoring procedure for NMT.

We conduct the experiments on three translation tasks: Uyghur→Chinese, Chinese→English and English→French, which represent low-resource, medium-resource and high-resource NMT, respectively. The experimental results show the proposed approaches achieve comparable performances with empirical length normalization methods. Further analyses show the flexibility of the proposed methods and the assumptions that our approaches rely on are reliable.

2 Related Work

The length bias reduction methods can be mainly divided into two categories: i) dividing the log probability by the length penalty lp:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}})/lp, \tag{4}$$

and ii) adding an additive length-related reward to the log probability of the hypothesis:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) + \gamma \cdot r. \tag{5}$$

For the first branch, the predominant form of the length penalty lp is the length of the hypothesis (Boulanger-Lewandowski et al., 2013; Jean et al., 2015; Koehn and Knowles, 2017; Meister et al., 2020). Google's NMT system (Wu et al., 2016) employ an empirical length penalty that is computed as:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) / \frac{(5 + T_{\hat{\mathbf{y}}})^{\alpha}}{(5+1)^{\alpha}}, \tag{6}$$

where the parameter α is used to control the strength of the length normalization. Stahlberg and Byrne (2019) apply another variant of lp, which introduces the information of the length ratio of the hypotheses over the source sentence. Yang et al. (2018) propose a brevity penalty normalization which adds the log brevity penalty bp to the normalized score:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}})/T_{\hat{\mathbf{y}}} + \log bp, \tag{7}$$

where bp is same as the form of brevity penalty in BLEU (Papineni et al., 2002):

$$bp = \begin{cases} 1 & gr \cdot T_{\mathbf{x}} < T_{\hat{\mathbf{y}}} \\ e^{(1 - T_{\mathbf{y}}/T_{\hat{\mathbf{y}}})} & gr \cdot T_{\mathbf{x}} \geq T_{\hat{\mathbf{y}}} \end{cases}, \tag{8}$$

where gr is the generation ratio i.e. $T_{\mathbf{y}}/T_{\mathbf{x}}$. Since $T_{\mathbf{y}}$ is unknown in the inference step, Yang et al. (2018) apply a 2-layer multi layer perceptron (MLP) to predict the gr by taking the mean of the hidden states of the NMT encoder as the input.

The second branch is similar to the word penalty in statistical machine translation (Och and Ney, 2002; Koehn, 2010). The parameter γ can be automatically optimized with supervised learning (He et al., 2016; Murray and Chiang, 2018) or manually assignment (Huang et al., 2017).

He et al. (2016) propose a log-linear NMT framework which incorporates a word reward feature to the framework to control the length of the translation:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) + \gamma \cdot T_{\hat{\mathbf{y}}}, \tag{9}$$

where γ is trained with other parameters of the log-linear NMT model using minimum error rate training (Och, 2003; He et al., 2016). Murray and Chiang (2018) make the optimization process of γ independent to the NMT training process, so that the γ can be trained on a relatively small dataset. Huang et al. (2017) introduce a Bounded Length Reward that includes the prior knowledge of the generation ratio gr of reference translation length over source sentence length:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) + \gamma \cdot \min(gr \cdot T_{\mathbf{x}}, T_{\hat{\mathbf{y}}}), \tag{10}$$

where the length reward γ is fine-tuned manually. All the above methods (He et al., 2016; Murray and Chiang, 2018; Huang et al., 2017) fine-tune the correcting ratio γ by a supervised data, which may lead to less optimal results on unseen test datasets. Yang et al. (2018) propose a Bounded Adaptive-Reward to remove the hyperparameter γ: $s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) + \sum_{t=1}^{T^*} r_t$, where b is the beam size and r_t is the average negative log-probability of the words in the beam at time step t. $T^* = \min\{T_{\hat{\mathbf{y}}}, T_{pred}(x)\}$, where $T_{pred}(x)$ is predicted with a 2-layer MLP instead of using the constant gr (Huang et al., 2017) as Eq. (10) does.

The proposed HSR-based debiasing method is motivated entirely by a causal structure shown in Fig. 1, although the form of the approach is same as the reward-based length normalization in Eq. (5). Formally, we can regard $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ in Eq. (3) as an instance of $(\gamma \cdot r)$ in Eq. (9) with very few prior assumptions or handcrafted designs. The leaning process of $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ is entirely model agnostic and unsupervised, which makes the proposed method more competitive to the previous supervised approaches (He et al., 2016; Murray and Chiang, 2018; Huang et al., 2017) in real practical applications.

3 Approach

3.1 Correcting Length Bias via Half-Sibling Regression

In this paper, we apply a debiasing framework of Half-Sibling Regression (HSR) (Schölkopf et al., 2016) to subtract the NMT scoring bias caused by the length of the translation. For a translation hypothesis

Algorithm 1 HSR in translation re-scoring for correcting length bias. See section 3.2 for more details.

```
Input: $m$ translation candidates: $\hat{\mathrm{Y}} = \{\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_m\}$, the lengths set of the translation candidates: $T(\hat{\mathrm{Y}}) = \{T_{\hat{\mathbf{y}}_1}, \cdots, T_{\hat{\mathbf{y}}_m}\}$, NMT model scores for the $m$ translation candidates: $s(\hat{\mathrm{Y}}) = \{s(\hat{\mathbf{y}}_1), \cdots, s(\hat{\mathbf{y}}_m)\}$ and a hyperparameter $\alpha \in [0, 1]$ .
Find the optimal parameters $\theta_R^*$ for a regression model $\mathrm{R}(T_{\hat{\mathbf{y}}}; \theta_R)$ by minimize the mean square error:
    $$\theta_R^* = \arg\min_{\theta_R} \frac{1}{m} \sum_{\hat{\mathbf{y}} \in \hat{\mathrm{Y}}} |\mathrm{R}(T_{\hat{\mathbf{y}}}; \theta_R) - s(\hat{\mathbf{y}})|^2$$
Subtract length information from the model estimated score:
    $$s'(\hat{\mathrm{Y}}) \leftarrow s(\hat{\mathrm{Y}}) - \alpha \times \mathrm{R}^*(T(\hat{\mathrm{Y}}); \theta_R^*) \tag{12}$$
Output: The debiased translation scores $s'(\hat{\mathrm{Y}}) = \{s'(\hat{\mathbf{y}}_1), \cdots, s'(\hat{\mathbf{y}}_m)\}$.
```

$\hat{\mathbf{y}}$, suppose that Q is the true translation quality that we cannot observe directly, and we regard S as an observable degraded version of Q which is affected by Q and some noise N, simultaneously. Considering a conventional NMT decoder, S is usually calculated by $s(\hat{\mathbf{y}})$ in Eq. (2). As discussed in section 1, $T_{\hat{\mathbf{y}}}$, as the length of $\hat{\mathbf{y}}$, has undesired crucial impacts on S. We refer $s(\hat{\mathbf{y}})$ as a measurement of Q with systemic errors N, then $T_{\hat{\mathbf{y}}}$ is the correlative variable of N that satisfies $N \not\perp\!\!\!\perp T_{\hat{\mathbf{y}}}$. At the same time, we assume that $Q \perp\!\!\!\perp T_{\hat{\mathbf{y}}}$, therefore, we can subtract the effects of $T_{\hat{\mathbf{y}}}$ on S, i.e. $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$, from S to eliminate length bias without affect the connection between S and Q:

$$S' \leftarrow S - \mathrm{E}[S|T_{\hat{\mathbf{y}}}]. \tag{11}$$

In practice, the value of $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ can be estimated by a regression model that is trained on the observed $(S, T_{\hat{\mathbf{y}}})$ pairs.

Fig. 1 shows the causal directed acyclic graph (DAG) that illustrates the causalities between Q, S, N, N^* and $T_{\hat{\mathbf{y}}}$, where N^* is other noise source that satisfies $N^* \perp\!\!\!\perp T_{\hat{\mathbf{y}}}$. We set up an undirected connection between N and $T_{\hat{\mathbf{y}}}$ to represent $N \not\perp\!\!\!\perp T_{\hat{\mathbf{y}}}$ since the causal direction between the two variables is not important in this paper. It is worth noting that $Q \perp\!\!\!\perp T_{\hat{\mathbf{y}}}$ is a strong assumption when we don't know the specific form of Q. The possible forms of Q and the assumption of $Q \perp\!\!\!\perp T_{\hat{\mathbf{y}}}$ will be discussed in more detail in section 3.3.1.

3.2 Re-scoring Translation Candidates

The HSR-based length debiasing method is model agnostic and it views the NMT model as a blackbox. Therefore, we simply apply the HSR-based approach to the translation re-scoring process to verify its effectiveness. Algorithm 1 shows a sketch of the proposed re-scoring framework. As described in Algorithm 1, we first optimize a regression model $\mathrm{R}(T_{\hat{\mathbf{y}}}; \theta_R)$ that parameterized by θ_R to estimate the length effect on $s(\hat{\mathbf{y}})$ by using the data $(T(\hat{\mathrm{Y}}), s(T_{\hat{\mathbf{y}}})) = \{(T_{\hat{\mathbf{y}}_i}, s(\hat{\mathbf{y}}_i))\}_{i=1}^m$. Then, we adopt the optimal $\mathrm{R}^*(T_{\hat{\mathbf{y}}}; \theta_R^*)$ as an approximate to $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ in Eq. (11) to eliminate the length information from $s(\hat{\mathbf{y}})$:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) - \alpha \times \mathrm{R}^*(T_{\hat{\mathbf{y}}}; \theta_R^*). \tag{13}$$

Following Wu et al. (2016), we introduce a hyperparameter $\alpha \in [0, 1]$ to control the strength of the debiasing operation. $\alpha = 0$ means no debiasing operation is conducted and empirical studies show that setting $\alpha = 1$ usually gains better performances for $b \geq 8$. (Note that Eq. (12) in Algorithm 1 is in a set form while Eq. (13) is in a single value form.)

We propose two branches of implementations for the proposed re-scoring framework in practice: i) a corpus based re-scoring method (C-HSR) and ii) a single source sentence based re-scoring method (S-HSR). For C-HSR, we perform the regression over the translations and their model scores of the whole test dataset, in other words, it needs the NMT model to finish translating the whole test set. For S-HSR, the regression model is optimized on the the translation candidates and their model scores of a single input source sentence. Therefore, the size of $\hat{\mathrm{Y}}$ in Algorithm 1, i.e. m, equals the beam size b and $b \times |\mathrm{X}_{test}|$ (the size of test set) for S-HSR and C-HSR, respectively.

3.3 Discussion

3.3.1 The Assumption of Q is Independent of $T_{\mathbf{y}}$

Considering one of ideally forms of Q that is straightforward defined as a conditional probability:

$$Q := P(\mathbf{y}|\mathbf{x}) = P(\{y_1, ..., y_{T_{\mathbf{y}}}\}|\mathbf{x}). \tag{14}$$

In Eq. (14), $T_{\hat{\mathbf{y}}}$ is an inherent feature of $\mathbf{y}$, so it is also involved in Q. Therefore, executing the calculation of Eq. (11) will inevitably eliminate parts of Q itself.

However, the condition where $T_{\mathbf{y}}$ is almost independent of Q is also sufficient for HSR in practice, according to Schölkopf et al. (2016). Hence, we should verify the correlation between Q and $T_{\hat{\mathbf{y}}}$ before employing our approach to specific applications. Since, Q as well as $P(\mathbf{y}|\mathbf{x})$ is theoretic and unobservable, we adopt a more precise and pricey observable variable, the professional translators' direct assessment (DA) score, as an approximation to the Q [1]. We use the datasets from WMT 2020 Quality Estimation Share Task 1[2]: Sentence-Level Direct Assessment (Specia et al., 2020) to analyze the Pearson's and Spearman's correlation scores between the length of translation and the DA score, and the results are presented in Table 1.

As Table 1 shows, for most conditions, the absolute values of the correlation scores are less than 0.20, which indicates that Q is almost independent of the translation length in a linear 2-dimensional space. However, there are multiple possible variables that influence the human DA score such as the number of the rare words in the source sentence and the translation hypothesis. Although partial correlation (Baba et al., 2004) might be effective for analyzing multiple correlative variables, the information about the other observable variables is unavailable. In general, we believe that removing $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ will not harm the information of Q too much, and the debiasing ratio α is also a conservative design to avoid punishing the length information overly.

Language Pair	Train		Valid		Test	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
English-German	-0.06	-0.11	-0.15	-0.18	-0.18	-0.18
English-Chinese	-0.07	-0.12	-0.08	-0.09	-0.00	-0.02
Romanian-English	-0.20	-0.15	-0.20	-0.14	-0.25	-0.18
Estonian-English	-0.09	-0.13	-0.09	-0.10	-0.11	-0.11
Nepalese-English	-0.12	-0.02	-0.12	-0.05	-0.09	-0.01
Sinhala-English	-0.14	-0.06	-0.11	-0.05	-0.17	-0.07
Russian-English	0.07	-0.07	0.00	-0.10	-0.01	-0.16

Table 1: The Pearson's and Spearman's correlation scores between the DA score and $T_{\hat{\mathbf{y}}}$.

3.3.2 The Connection to the Word Reward

The proposed HSR-based debiasing method is motivated by a causal structure, although the formalized form of our proposed approach is same as adding length-related reward in Eq. (5), by regarding $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ as a special instance of $(\gamma \cdot r)$. In particular, if we only consider the linear effects, i.e. $\mathrm{R}(T_{\hat{\mathbf{y}}}; \theta_R) = \theta_1 T_{\hat{\mathbf{y}}} + \theta_2$, then Eq. (13) is expand as:

$$s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}}) - \alpha \times (\theta_1^* T_{\hat{\mathbf{y}}} + \theta_2^*) = s(\hat{\mathbf{y}}) - \alpha\theta_1^* T_{\hat{\mathbf{y}}} - \alpha\theta_2^*, \tag{15}$$

which is similar to the word reward in Eq. (9). The $\theta_1^* \in \mathbb{R}$ and $\theta_2^* \in \mathbb{R}$ in Eq. (15) are optimal parameters of the linear regression. Therefore, under the above linear assumption, the proposed method can be seen as a simple and effective unsupervised strategy to optimize γ for the word penalty (He et al.,

[1] Note that, $P(\mathbf{y}|\mathbf{x})$ is one of the formal definitions of Q, and it is not the essence of Q. On the other hand, the human generated DA score is the currently available best approximation of Q to our best knowledge.

[2] http://www.statmt.org/wmt20/quality-estimation-task.html

2016; Murray and Chiang, 2018). Since most of the previous word penalty efforts determine γ through a supervised procedure (He et al., 2016; Murray and Chiang, 2018; Huang et al., 2017) before the testing stage, they may fall into less optimal results on unseen datasets.

However, if we do not apply the linear regression, the form will be different to the word penalty. In this paper, we study the performances of various typical regression models including linear regression, support vector regression, k-neighbors regression, multi-layer perceptron (MLP) regression and random forest regression. We find that applying linear regression and MLP regression to C-HSR and S-HSR respectively gain better performances. The detailed analyses about various regression models are shown in section 5.1.

4 Experiments

4.1 Datasets and Evaluation Metric

We evaluate the proposed approaches on three translation tasks: Uyghur→Chinese (Ug→Zh), Chinese→English (Zh→En) and English→French (En→Fr). For each of the translation task, the corpus is tokenized by the Moses (Koehn et al., 2007) *tokenizer.perl* [3] before encoded with byte-pair encoding (Sennrich et al., 2016). For Zh→En and Ug→Zh translation tasks, the Chinese parts are segmented by the LTP segmentor (Che et al., 2010) before tokenizing.

Ug→Zh. For Uyghur→Chinese translation, the training corpus is from Uyghur to Chinese News Translation Task in CCMT2019 Machine Translation Evaluation (Yang et al., 2019). Apart from the Moses (Koehn et al., 2007) tokenizer, we do not use any other tools to segment Uyghur. The training set contains 0.17M parallel sentence pairs, and the vocabularies are 30K for both Uyghur and Chinese corpus. The official validation set and the test set are applied in our experiments.

Zh→En. For Chinese→English translation, the training data is extracted from four LDC corpora[4]. The training set finally contains 1.3M parallel sentence pairs in total. After preprocessing, we get a Chinese vocabulary of about 39K tokens, and an English vocabulary of about 30K tokens. We use NIST2002 dataset for validation and NIST 2003∼2006 datasets for test.

En→Fr. For English→French translation, we conduct our experiments on the publicly available WMT'14 En→Fr datasets which consist of 18M sentences pairs. Both source and target vocabulary contains 30K tokens after preprocessing. We report results on newstest2014 dataset, and newstest2013 dataset is used as the validation set.

Evaluation. Following Vaswani et al. (2017), we report the results of a single model by averaging the 5 checkpoints around the best model selected on the development set. The translation results are measured in case-insensitive BLEU (Papineni et al., 2002) by *multi-bleu.perl*[3]. For the Ug→Zh translation task, the BLEU scores are reported at character-level.

4.2 Length Normalization Baselines

We adopt two popular empirical length normalization strategies ((i),(ii)) and a complicated MLP-based method ((iii)) as the comparison baseline methods: i) Length Norm: directly dividing the translation score by the length of the translation (Boulanger-Lewandowski et al., 2013; Jean et al., 2015; Koehn and Knowles, 2017) as shown in Eq. (4), ii) GNMT: the length normalization method of Google NMT (Wu et al., 2016), as shown in Eq. (6), and iii) BP Norm: the length normalization method that applies a model predicted *bp* constraint (Yang et al., 2018) as shown in Eq. (7) and Eq. (8). We average the outputs of the Transformer encoder instead of the LSTM hidden layers as the input of the 2-layers MLP used in Yang et al. (2018). For fairness considerations, those methods are all unsupervised[5], since our proposed methods do not rely on any human reference.

[3] Moses scripts: `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/`

[4] LDC2005T10, LDC2003E14, LDC2004T08 and LDC2002E18. Since LDC2003E14 is a document-level alignment comparable corpus, we use Champollion Tool Kit (Ma, 2006) to extract parallel sentence pairs from it.

[5] "unsupervised" means that the method is not trained on the dataset that consists the pairs of translation hypothesis and human reference.

4.3 Model Setups

We apply the base model of Transformer (Vaswani et al., 2017) as the specific implement of the NMT baseline in our work, and we build up the NMT models based on OpenNMT-py (Klein et al., 2017).

We analyze different regression models for both C-HSR and S-HSR, and finally select linear regression for C-HSR and one-hidden layer MLP regression S-HSR, denoted by "C-HSR$_{LR}$" and "S-HSR$_{MLP}$", respectively. The regression models used in our work are implemented by using scikit-learn (Pedregosa et al., 2011). The detailed setups and analyses for different regression models are shown in section 5.1.

Following Wu et al. (2016), we use α to control the strength of length bias correcting. The α is selected according to the performance on the validation set and the detail selections of α for different model setups are shown in Table 2. The detailed analyses on α are presented in section 5.2.

Language Pair	Method	b=4	b=8	b=16	b=32	b=64	b=100	b=200
Ug→Zh	*GNMT*	1.0	-	-	-	-	-	-
	C-HSR$_{LR}$	0.9	1.0	-	-	-	-	-
	S-HSR$_{MLP}$	1.0	-	-	-	-	-	-
Zh→En	*GNMT*	0.5	0.9	1.0	-	-	-	-
	C-HSR$_{LR}$	0.7	1.0	-	-	-	-	-
	S-HSR$_{MLP}$	0.7	0.9	0.9	0.9	1.0	-	-
En→Fr	*GNMT*	0.9	-	-	-	-	-	-
	C-HSR$_{LR}$	0.8	-	-	-	-	0.9	1.0
	S-HSR$_{MLP}$	0.8	-	-	-	-	-	-

Table 2: **Correcting ratio α for different model setups**. "-" means same as the value in the left cell.

4.4 Main Results

We conduct experiments on three translation tasks with disparate corpora scales: low-resource Ug→Zh, medium-resource Zh→En and high-resource En→Fr. We present BLEU scores on translations with two different decoding beam sizes: $b = 4$ and $b = 200$, in order to compare the model performances on small and large beam sizes. The experimental results are shown in Table 3 and Table 4.

The overall results show that all the length debiasing approaches obtain better BLEU scores than the baseline NMT model for large beam size. For the condition of smaller beam size, "*Length Norm*" tends to disrupt the model performance on En→Fr and Zh→En datasets, which is contrary to the case of a larger search space.

Our proposed C-HSR$_{LR}$ and S-HSR$_{MLP}$ seem to produce stable BLEU scores across multiple datasets and beam sizes. The results show that S-HSR$_{MLP}$ usually gains better BLEU scores than C-HSR$_{LR}$ on the large beam size (Ug→Zh, Zh→En and En→Fr), while C-HSR$_{LR}$ performs better on the small beam size (Zh→En and En→Fr). We consider the reason is that S-HSR$_{MLP}$ is trained better on $b = 200$ than that on small dataset. On the other hand, the requirements for training a linear regression model is not as strict as it for MLP, although the accuracy of the linear model may be lower than the MLP-based model when both of them are well trained.

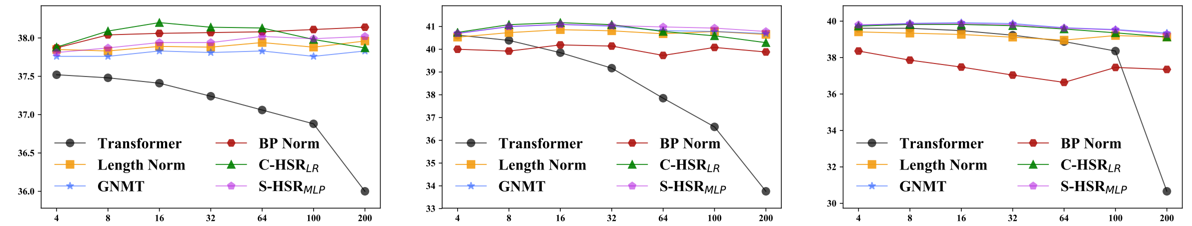
The performance of BP Norm is unsatisfactory, which we consider the reason is that the MLP-based generation ratio predictor does not work well. If our hypothesis is correct, the length of the transla-

Method	En→Fr		Ug→Zh	
	b=4	b=200	b=4	b=200
Transformer	39.61	30.66	37.52	36.00
+*Length Norm*	39.41	39.13	37.85	37.96
+*GNMT*	39.77	**39.35**	37.76	37.83
+*BP Norm*	38.36	37.35	37.87	**38.14**
+C-HSR$_{LR}$	39.73	39.13	**37.88**	37.87
+S-HSR$_{MLP}$	**39.80**	39.28	37.81	38.02

Table 3: **BLEU scores on En→Fr and Ug→Zh translation tasks**. "b" represents the beam size.

Method	03		04		05		06		Average	
	b=4	=200	b=4	=200	b=4	=200	b=4	=200	b=4	=200
Transformer	40.10	33.55	42.09	35.31	40.33	33.46	39.94	32.71	40.62	33.76
+*Length Norm*	39.99	40.13	42.05	42.23	39.67	40.10	**40.42**	**40.14**	40.53	40.65
+*GNMT*	40.13	40.08	42.18	42.18	**40.39**	**40.59**	40.24	39.89	40.74	40.69
+*BP Norm*	39.46	39.25	41.50	41.22	39.19	39.15	39.84	39.91	40.00	39.88
+C-HSR$_{LR}$	40.35	39.58	**42.60**	42.00	40.32	40.22	40.25	39.34	**40.88**	40.29
+S-HSR$_{MLP}$	**40.40**	**40.25**	42.42	**42.44**	40.33	40.40	40.25	40.04	40.85	**40.78**

Table 4: **BLEU scores on NIST 2003∼2006 Zh→En translation task**. “b” represents the beam size.



(a) Ug→Zh

(b) Zh→En

(c) En→Fr

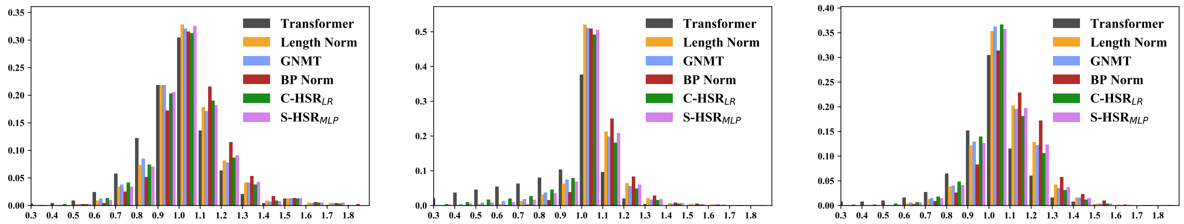
Figure 2: **BLEU scores of different methods with respect to different beam sizes** $[4 \sim 200]$**.** The y-axis is the BLEU score, and the x-axis is the decoding beam size. For Zh→En task, we present the averaged the BLEU score of NIST 2003∼2006. See section 4.5 for more details.

tion will be too long or too short under the rescore method of BP Norm. Further analyses about the performances of those method on various beam sizes are shown in section 4.5.

4.5 Performance on Wider Beam Size

As a supplement to section 4.4, we analyze the performances of the proposed approaches on different decoding beam sizes. Fig. 2 shows the trend of the BLEU scores with respect to the beam sizes of $[4, 8, 16, 32, 64, 100, 200]$ for the three translation tasks. From Fig. 2 we can observe that all the length debiasing methods achieve stable and comparable performances when the beam size increases.

We also analyze the length ratio distributions between generated sentences and references $(T_{\hat{\mathbf{y}}}/T_{\mathbf{y}})$, and Fig. 3 shows the length ratio on the decoding beam size of 200 for the three translation tasks. The histogram illustrates that the Transformer baseline select more short translations with $b = 200$, while all the debiasing methods avoid the length problem effectively. We can see that BP Norm selects most translations that are longer than the human references, which explains their deficient performances on Zh→En and En→Fr translation tasks.



(a) Ug→Zh

(b) Zh→En

(c) En→Fr

Figure 3: **Histogram of the distributions of length ratio between the generated sentences and the human references across methods with the beam size of** 200**.** The y-axis is the frequency, and the x-axis is the length ratio. See section 4.5 for more details.

5 Analysis

5.1 Comparisons among Different Regression Models

The proposed HSR-based debiasing framework has a large elbow room for the implementation of $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$. In this paper, we analyze 5 typical regression methods for making the estimation of $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$ to verify the flexibility of the proposed method. The models and setups are listed in Table 5.

#	Method	Regression (Abbr)	Parameters Setups
1	C-HSR	Linear (LR)	<default>
		Support Vector (SV)	Not used in C-HSR
2		K Neighbors (KN)	n_neighbors=2, weights=”distance”
3		Multi-layer Perceptron (MLP)	hidden_layer_sizes=50, activation=’relu’, max_iter=10
4		Random Forest (RF)	n_estimators=9, criterion=’mse’
5	S-HSR	Linear (LR)	<default>
6		Support Vector (SV)	kernel=’rbf’, C=100, tol=1.5, epsilon=0.1, gamma=’auto’
7		K Neighbors (KN)	n_neighbors=2, weights=”distance”
8		Multi-layer Perceptron (MLP)	hidden_layer_sizes=50, activation=’relu’, max_iter=35
9		Random Forest (RF)	n_estimators=3, criterion=’mse’

Table 5: **Model setups for different regression methods.** “Regression (Abbr)” is the regression model and its abbreviation name in this paper. “Parameters Setup” is the hyper-parameters set for the model in scikit-learn toolkits (Pedregosa et al., 2011), and other parameters that are not mentioned are set by default values.

The model setups shown in Table 5 are fine-tuned on Zh→En NIST 2002 dataset with the beam size of 200. We apply the same model setups for all the datasets and beam sizes, in order to verify the generalization of the proposed HSR-based debiasing method, though the optimal hyper-parameters of the regression models may be distinct. We do not apply SV to C-HSR, since we find that the support vector tends to be a horizontal line in this work when the size of training set is larger than $1,000$.

The BLEU scores and the executing speed on the three translation tasks are shown in Table 6. The operating environment is CentOS Linux release 7.5.1804 with Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz. The overall results show that the S-HSR methods (#5∼#9) usually gains better BLEU scores than the C-HSR methods (#1∼#4) on large beam size no matter which regression model is used. It illustrates that the S-HSR methods are more adaptable to different translation hypotheses than the C-HSR methods.

On the other hand, different regression models have little differences in BLEU scores, and each regression method obtains comparable results with other regression models. It illustrates that the principles and the assumptions that the proposed methods rely on are reliable, since the specific implements of the approaches do not play an important role on the experimental results. However, from the aspect of executing speed, applying linear regression shows a remarkable advantage comparing with the others.

5.2 Selection of the Correcting Ratio α

As Table 2 shows, the optimal selections of α are usually distinctive for different datasets and beam sizes. We study the impact of different α on the BLEU scores by taking NIST 2002 Zh→En development set as an example, and the experimental results are shown in Table 7. The overall results show that for a small search space, an $\alpha \in [0.5, 0.9]$ is appropriate, while for larger beam sizes, setting $\alpha = 1.0$ usually gets better BLEU scores. There may be two reasons for the above phenomenon: i) the model generated score $s(\hat{\mathbf{y}})$ is more severely affected by $T_{\hat{\mathbf{y}}}$ with the large decoding beam size, and ii) large beam size provides the regression model with more training data to estimate $\mathrm{E}[S|T_{\hat{\mathbf{y}}}]$. We compare the results among C-HSR$_{LR}$, C-HSR$_{MLP}$ and S-HSR$_{MLP}$ to control the impacts of different regression models.

The size of the training data is $b \times |\mathrm{X}_{nist02}|$ for the C-HSR methods, which is sufficient on each beam size. Therefore, we consider that the $s(\hat{\mathbf{y}})$ is less affected by $T_{\hat{\mathbf{y}}}$ with small beam size, since the optimal α is 0.7 for both C-HSR and S-HSR. On the other hand, comparing with the BLEU scores under $b = 32$ and $b = 200$, we believe that the size of training data is another factor to influence the selection of α, especially for S-HSR.

#	Method	Ug→Zh		Zh→En		En→Fr		Speed (sent./s)	
		b=4	b=200	b=4	b=200	b=4	b=200	b=4	b=200
1	C-HSR$_{LR}$	**37.88**	37.87	**40.88**	40.29	39.73	39.13	7,934	163
2	C-HSR$_{KN}$	37.73	37.16	40.41	40.00	39.05	37.97	781	13
3	C-HSR$_{MLP}$	37.85	37.19	40.89	40.16	39.61	37.94	2,556	60
4	C-HSR$_{RF}$	37.75	37.36	40.76	40.09	39.65	37.83	475	9
5	S-HSR$_{LR}$	37.77	37.87	40.67	40.24	39.69	39.02	2,084	117
6	S-HSR$_{SV}$	37.57	37.75	40.60	40.60	39.60	38.90	1,930	61
7	S-HSR$_{KN}$	37.65	38.00	40.67	40.66	39.63	38.51	779	14
8	S-HSR$_{MLP}$	37.81	**38.02**	40.85	**40.78**	**39.80**	**39.28**	84	12
9	S-HSR$_{RF}$	37.62	37.78	40.62	40.63	39.69	38.69	157	15

Table 6: **BLEU and executing speed comparison among different setups of our approaches.** For Zh→En task, we present the averaged the BLEU score of NIST 2003∼2006. "Speed" is measured in sentences per second (sent./s). See section 5.1 for more details.

b	Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
4	C-HSR$_{LR}$	41.81	41.87	41.91	41.91	41.96	41.92	**42.01**	41.98	41.91	41.84
	C-HSR$_{MLP}$	41.81	41.87	41.93	41.90	41.95	41.98	**41.98**	41.92	41.84	41.78
	S-HSR$_{MLP}$	41.81	41.87	41.93	41.98	41.98	41.93	**41.99**	41.93	41.90	41.77
32	C-HSR$_{LR}$	40.59	40.56	40.68	41.13	41.14	41.22	41.54	41.84	41.99	**42.20**
	C-HSR$_{MLP}$	40.50	40.59	40.67	41.00	41.19	41.27	41.61	41.80	41.95	**42.11**
	S-HSR$_{MLP}$	40.52	40.61	40.88	40.94	41.08	41.16	41.50	41.78	**42.02**	41.95
200	C-HSR$_{LR}$	34.41	35.22	36.09	37.06	38.41	39.18	39.87	40.60	41.23	**41.57**
	C-HSR$_{MLP}$	34.25	34.95	35.65	36.38	37.46	38.60	39.55	40.40	40.86	**41.49**
	S-HSR$_{MLP}$	34.36	35.71	36.48	37.67	38.65	39.92	40.86	41.35	41.72	**42.07**

Table 7: **BLEU score with respect to different values of α on NIST 2002 Zh→En dataset.** "b" represents the beam size. "0.1 ∼ 1.0" are the values of α. See section 5.2 for more details.

In this paper, the introduction of the correcting ratio α can slack the strict assumption of $Q \perp\!\!\!\perp S$, which prevent the proposed approaches from punishing length information too much. The results presented in Table 2 and Table 7 verify parts of our discussion about the independent assumptions.

6 Conclusion and Future Work

In this paper, we introduce a causal motivated method to reduce the length bias problem in NMT. We employ a Half-Sibling Regression (Schölkopf et al., 2016) method to handle this task and corroborate the task satisfies the independence assumption of HSR. Experimental results on three language pairs with distinct data scales show the effectiveness of the proposed method.

In the future, we will complete our experiments on the task of Quality Estimation. Since the proposed approaches are model agnostic and unsupervised, we will verify the effectiveness of our approaches on other natural language generation tasks, such as dialogue system and summarization. We also hope to make further studies about the relationships among Q, S and $T_{\mathbf{y}}$.

Acknowledgments

We thank all anonymous reviewers for their valuable comments. This work is supported by the National Key Research and Development Program of China (Grant No. 2017YFB1002103) and the National Natural Science Foundation of China (No. 61732005).

References

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. 2004. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664.

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

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. 2013. Audio chord recognition with recurrent neural networks. In Alceu de Souza Britto Jr., Fabien Gouyon, and Simon Dixon, editors, *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, pages 335–340.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. LTP: A Chinese language technology platform. In *Coling 2010: Demonstrations*, pages 13–16, Beijing, China, August. Coling 2010 Organizing Committee.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with SMT features. In Dale Schuurmans and Michael P. Wellman, editors, *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 151–157. AAAI Press.

Liang Huang, Kai Zhao, and Mingbo Ma. 2017. When to finish? optimal beam search for neural text generation (modulo beam size). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2134–2139, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. 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, June. Association for Computational Linguistics.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *CoRR*, abs/1601.00372.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online, November. Association for Computational Linguistics.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium, October. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. 2016. Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. USA*, 113(27):7391–7398.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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, August. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online, November. Association for Computational Linguistics.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? 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 3356–3362, Hong Kong, China, November. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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 Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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. *CoRR*, abs/1609.08144.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium, October-November. Association for Computational Linguistics.

Muyun Yang, Xixin Hu, Hao Xiong, Jiayi Wang, Yiliyaer Jiaermuhamaiti, Zhongjun He, Weihua Luo, and Shujian Huang. 2019. Ccmt 2019 machine translation evaluation report. In Shujian Huang and Kevin Knight, editors, *Machine Translation*, pages 105–128, Singapore. Springer Singapore.