Check for updates

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(✉)], and Yi-Kun Tang[1,2]

[1] School of Computer Science and Technology, Beijing Institute of Technology,
Beijing 100081, China

[2] Beijing Engineering Research Center of High Volume Language Information
Processing and Cloud Computing Applications, Beijing 100081, China

{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.

Keywords: Machine translation · Causal inference · Half-sibling regression

1 Introduction

Recently, with the renaissance of deep learning, end-to-end neural machine translation (NMT) [2,26] has gained remarkable performances [6,27,28]. NMT models are usually built upon an encoder-decoder framework [5]: 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:

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

Most existing NMT approaches employ beam search to obtain more translation candidates and then gain a better translation hypothesis $\hat{\mathbf{y}} = \{\hat{y}_1, \cdots, \hat{y}_{T_{\hat{\mathbf{y}}}}\}$

© Springer Nature Switzerland AG 2021
S. Li et al. (Eds.): CCL 2021, LNAI 12869, pp. 3–15, 2021.
https://doi.org/10.1007/978-3-030-84186-7_1

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 [13]. 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. $s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}})/lp$ [3,9,13,16,30], and ii) adding an additional length-related reward r to $s(\hat{\mathbf{y}})$, i.e. $s'(\hat{\mathbf{y}}) \leftarrow s(\hat{\mathbf{y}})+\gamma\cdot r$ [7,8,14,17,30]. For the second strategy, the correcting ratio γ of the reward is usually determined by supervised training [7,17] or manually fine-tuning [8] 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) [22] 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.

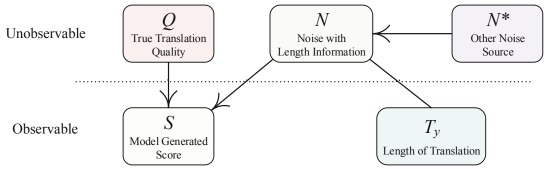


Fig. 1. A causal directed acyclic graph shows the relations among the true translation quality Q, the model generated score S and the translation length $T_{\hat{\mathbf{y}}}$. See Sect. 1 and Sect. 3.1 for more details.

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 [3,9,13,16]. Google's NMT system [28] 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 [25] apply another variant of lp, which introduces the information of the length ratio of the hypotheses over the source sentence. Yang et al. [30] 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 [20]:

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

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. [30] 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 [11,19]. The parameter γ can be automatically optimized with supervised learning [7,17] or manually assignment [8].

He et al. [7] 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 [7,18]. Murray and Chiang [17] 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. [8] 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 [7,8,17] fine-tune the correcting ratio γ by a supervised data, which may lead to less optimal results on unseen test datasets. Yang et al. [30] 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 [8] 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 [7,8,17] 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) [22] to subtract the NMT scoring bias caused by the length of the translation. For a translation hypothesis $\hat{\mathbf{y}}$, 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 Sect. 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

Algorithm 1 . HSR in translation re-scoring for correcting length bias. See Sect. 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 $\alpha \in [0, 1]$ .
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:
       $$\theta_R^* = \arg\min_{\theta_R} \frac{1}{m} \sum_{\hat{\mathbf{y}} \in \hat{\mathrm{Y}}} |\mathrm{R}(T_{\hat{\mathbf{y}}}; \theta_R) - s(\hat{\mathbf{y}})|^2$$
2: Subtract length information from the model estimated score:
       $$s'(\hat{\mathrm{Y}}) \leftarrow s(\hat{\mathrm{Y}}) - \alpha \times \mathrm{R}^*(T(\hat{\mathrm{Y}}); \theta_R^*) \qquad (12)$$
Output: The debiased translation scores $s'(\hat{\mathrm{Y}}) = \{s'(\hat{\mathbf{y}}_1), \cdots, s'(\hat{\mathbf{y}}_m)\}$.
```

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.

Figure 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 Sect. 3.3.

3.2 Re-scoring Translation Candidates

The HSR-based length debiasing method is model agnostic and it views the NMT model as a black-box. 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 [28], 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 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

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 [22]. 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 [24] 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 [1] 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.

[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.

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

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

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 [7,17]. Since most of the previous word penalty efforts determine γ through a supervised procedure [7,8,17] 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.

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 [12] *tokenizer.perl*[3] before encoded with byte-pair encoding [23]. For

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

Zh→En and Ug→Zh translation tasks, the Chinese parts are segmented by the LTP segmentor [4] before tokenizing.

Ug→Zh. For Uyghur→Chinese translation, the training corpus is from Uyghur to Chinese News Translation Task in CCMT2019 Machine Translation Evaluation [29]. Apart from the Moses [12] 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 [27], 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 [20] by *multi-bleu.perl* (see footnote 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 [3,9,13] as shown in Eq. (4), ii) GNMT: the length normalization method of Google NMT [28], as shown in Eq. (6), and iii) BP Norm: the length normalization method that applies a model predicted *bp* constraint [30] 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 [30]. For fairness considerations, those methods are all unsupervised[5], since our proposed methods do not rely on any human reference.

[4] LDC2005T10, LDC2003E14, LDC2004T08 and LDC2002E18. Since LDC2003E14 is a document-level alignment comparable corpus, we use Champollion Tool Kit [15] 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 [27] as the specific implement of the NMT baseline in our work, and we build up the NMT models based on OpenNMT-py [10]. 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 [21]. Following [28], 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.

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

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	–	–	–	–	–	–

4.4 Main Results

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

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 4. BLEU scores on NIST 2003∼2006 Zh→En translation task.

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**

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.

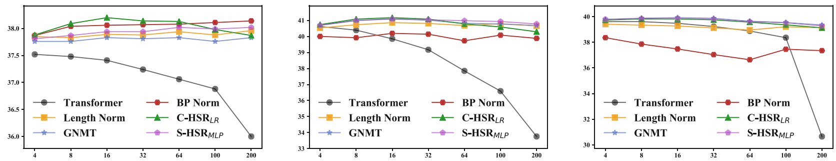
The overall results show that all the length debiasing approaches obtain better BLEU scores than the baseline NMT model for large beam size. For the condition of smaller beam size, "*Length Norm*" tends to disrupt the model performance on En→Fr and Zh→En datasets, which is contrary to the case of a larger 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 translation 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 Sect. 4.5.

4.5 Performance on Wider Beam Size

As a supplement to Sect. 4.4, we analyze the performances of the proposed approaches on different decoding beam sizes. Figure 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.



(a) Ug→Zh (b) Zh→En (c) En→Fr

Fig. 2. BLEU scores of different methods with respect to different beam sizes [4–200]. 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 Sect. 4.5 for more details.

5 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 [22] 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.

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

1. Baba, K., Shibata, R., Sibuya, M.: Partial correlation and conditional correlation as measures of conditional independence. Austr. New Zealand J. Stat. **46**(4), 657–664 (2004)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)
3. Boulanger-Lewandowski, N., Bengio, Y., Vincent, P.: Audio chord recognition with recurrent neural networks. In: de Souza Britto, A., Jr., Gouyon, F., Dixon, S. (eds.) Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, 4–8 November 2013, pp. 335–340 (2013)
4. Che, W., Li, Z., Liu, T.: LTP: a Chinese language technology platform. In: Coling 2010: Demonstrations, pp. 13–16. Coling 2010 Organizing Committee, Beijing, August 2010

5. Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y.: 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, pp. 103–111. Association for Computational Linguistics, Doha, October 2014
6. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 1243–1252. PMLR (2017)
7. He, W., He, Z., Wu, H., Wang, H.: Improved neural machine translation with SMT features. In: Schuurmans, D., Wellman, M.P. (eds.) Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, 12–17 February 2016, pp. 151–157. AAAI Press (2016)
8. Huang, L., Zhao, K., Ma, M.: 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, pp. 2134–2139. Association for Computational Linguistics, Copenhagen, September 2017
9. Jean, S., Firat, O., Cho, K., Memisevic, R., Bengio, Y.: Montreal neural machine translation systems for WMT 2015. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 134–140. Association for Computational Linguistics, Lisbon, September 2015
10. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.: OpenNMT: open-source toolkit for neural machine translation. In: Proceedings of ACL 2017, System Demonstrations, pp. 67–72. Association for Computational Linguistics, Vancouver, July 2017
11. Koehn, P.: Statistical Machine Translation. Cambridge University Press, New York (2010)
12. Koehn, P., et al.: 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, pp. 177–180. Association for Computational Linguistics, Prague, June 2007
13. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation, pp. 28–39. Association for Computational Linguistics, Vancouver, August 2017
14. Li, J., Jurafsky, D.: Mutual information and diverse decoding improve neural machine translation. CoRR abs/1601.00372 (2016)
15. Ma, X.: Champollion: a robust parallel text sentence aligner. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). European Language Resources Association (ELRA), Genoa, Italy, May 2006
16. Meister, C., Cotterell, R., Vieira, T.: If beam search is the answer, what was the question? In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2173–2185. Association for Computational Linguistics, Online, November 2020
17. Murray, K., Chiang, D.: Correcting length bias in neural machine translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 212–223. Association for Computational Linguistics, Brussels, October 2018
18. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 160–167. Association for Computational Linguistics, Sapporo, July 2003

19. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 295–302. Association for Computational Linguistics, Philadelphia, July 2002
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, July 2002
21. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
22. Schölkopf, B., et al.: Modeling confounding by half-sibling regression. Proc. Natl. Acad. Sci. U.S.A. **113**(27), 7391–7398 (2016)
23. Sennrich, R., Haddow, B., Birch, A.: 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), pp. 1715–1725. Association for Computational Linguistics, Berlin, August 2016
24. Specia, L., et al.: Findings of the WMT 2020 shared task on quality estimation. In: Proceedings of the Fifth Conference on Machine Translation, pp. 743–764. Association for Computational Linguistics, Online, November 2020
25. Stahlberg, F., Byrne, B.: 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), pp. 3356–3362. Association for Computational Linguistics, Hong Kong, November 2019
26. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, Montreal, Quebec, Canada, 8–13 December 2014, pp. 3104–3112 (2014)
27. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017, pp. 5998–6008 (2017)
28. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. CoRR abs/1609.08144 (2016)
29. Yang, M., et al.: CCMT 2019 machine translation evaluation report. In: Huang, S., Knight, K. (eds.) CCMT 2019. CCIS, vol. 1104, pp. 105–128. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1721-1_11
30. Yang, Y., Huang, L., Ma, M.: 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, pp. 3054–3059. Association for Computational Linguistics, Brussels, October–November 2018