Comparing BERT-based Reward Functions for Deep Reinforcement Learning in Machine Translation

Yuki Nakatani **Tomoyuki Kajiwara** **Takashi Ninomiya**
Graduate School of Science and Engineering, Ehime University, Japan
{nakatani@ai., kajiwara@, ninomiya@}cs.ehime-u.ac.jp

Abstract

In text generation tasks such as machine translation, models are generally trained using cross-entropy loss. However, mismatches between the loss function and the evaluation metric are often problematic. It is known that this problem can be addressed by direct optimization to the evaluation metric with reinforcement learning. In machine translation, previous studies have used BLEU to calculate rewards for reinforcement learning, but BLEU is not well correlated with human evaluation. In this study, we investigate the impact on machine translation quality through reinforcement learning based on metrics that are more highly correlated with human evaluation. Experimental results show that reinforcement learning with BERT-based rewards can improve various evaluation metrics.

1 Introduction

Sequence-to-sequence models based on deep learning, such as attention-based LSTM (Bahdanau et al., 2015; Luong et al., 2015) and Transformer (Vaswani et al., 2017), are capable of generating fluent sentences and have been used successfully in many text generation tasks, such as machine translation (Tan et al., 2020) and text simplification (Alva-Manchego et al., 2020). Most previous studies on text generation use cross-entropy loss between references and output sentences to train the models based on maximum likelihood estimation for each token. Differentiability of cross-entropy loss enables gradient-based estimation in a supervised learning framework, but it has a *Loss-Evaluation Mismatch* problem (Ranzato et al., 2016; Wiseman and Rush, 2016) in case of machine translation, where loss functions and evaluation metrics are not consistent, e.g., cross-entropy loss vs. BLEU (Papineni et al., 2002). That is, an output sentence that is semantically adequate may receive an unfairly low evaluation due to a superficial disagreement with the reference sentence.

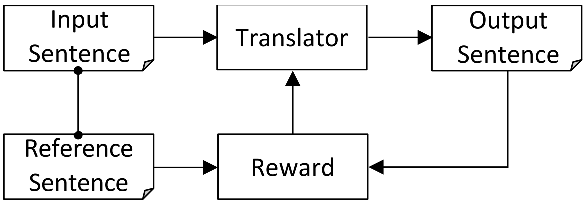

Figure 1: Machine translation based on deep reinforcement learning.

Such a Loss-Evaluation Mismatch problem (Ranzato et al., 2016; Wiseman and Rush, 2016) can be addressed by direct optimization of the evaluation metric through reinforcement learning (Williams, 1992). Since non-differentiable functions can be used as rewards in reinforcement learning, arbitrary evaluation metrics such as BLEU (Papineni et al., 2002), a word n-gram-based evaluation metric, and BLEURT (Sellam et al., 2020), an embedding-based evaluation metric, can be employed for the rewards of reinforcement learning. Performance improvements by using reinforcement learning have been reported in deep learning-based text generation, such as machine translation (Ranzato et al., 2016; Hashimoto and Tsuruoka, 2019; Yasui et al., 2019) and text simplification (Zhang and Lapata, 2017; Nakamachi et al., 2020).

In machine translation, many previous studies (Ranzato et al., 2016; Wu et al., 2018; Hashimoto and Tsuruoka, 2019; Kiegeland and Kreutzer, 2021) have used BLEU as rewards in reinforcement learning, but BLEU does not have a sufficiently high correlation with human evaluation. For machine translation metric tasks (Bojar et al., 2017), evaluation metrics have been proposed that correlate better with human evaluation than BLEU, such as chrF (Popović, 2017) and embedding-based evaluation metrics (Shimanaka et al., 2019; Zhang et al., 2020; Sellam et al., 2020) based on BERT (Devlin et al., 2019). Therefore, reward calculation using these evaluation metrics is expected to achieve further improvements in machine translation based on reinforcement learning.

This paper investigates the effectiveness of using surface-matching-based metrics and BERT-based metrics as the rewards for reinforcement learning in machine translation. Transformer-based machine translation models are trained in the reinforcement learning framework as shown in Figure 1. However, the action space for reinforcement learning of machine translation is very large because it deals with a vocabulary consisting of tens of thousands of tokens. Therefore, as in previous studies (Ranzato et al., 2016; Hashimoto and Tsuruoka, 2019), reinforcement learning is applied as fine-tuning to machine translation models that have been pre-trained by minimizing the cross-entropy loss. We then examine multiple metrics for both reward calculation and quality evaluation of machine translation, and investigate suitable reward functions for reinforcement learning of machine translation.

Experimental results on the IWSLT-2014 De-En translation task (Cettolo et al., 2014) revealed that reinforcement learning with BLEU as a reward function can only improve evaluation metrics based on surface matching, BLEU and chrF. On the other hand, reinforcement learning using BERT-based metrics as reward functions, such as BLEURT and BERT fine-tuned on the Semantic Textual Similarity (STS) estimation tasks (Cer et al., 2017), improved various metrics.

2 Reinforcement Learning for Machine Translation

In this study, pre-trained machine translation models are fine-tuned by deep reinforcement learning using various evaluation metrics as rewards. Section 2.1 describes pre-training of the machine translation model, followed by fine-tuning with reinforcement learning in Section 2.2, and finally, Section 2.3 outlines a machine translation metrics as a reward function for reinforcement learning.

2.1 Pre-training

The neural machine translation model consists of an encoder that encodes input sentences and a decoder that generates output sentences. The encoder is given a sequence of tokens of the input sentence $x = (x_0, x_1, ..., x_L)$ and outputs the hidden state $h = (h_0, h_1, ..., h_L)$. The decoder outputs the token sequence of the output sentence $y = (y_0, y_1, ..., y_M)$, given the hidden state h generated by the encoder. The probability of token y_t generation is maximized subject to x and $y_{<t} = (y_1, ..., y_{t-1})$. The log-likelihood of the output prediction is computed as follows.

$$\log p(y^i|x^i) = \sum_{t=1}^{M} \log p(y_t^i|y_{<t}^i, x^i) \qquad (1)$$

Pre-training minimizes the following cross-entropy loss for a dataset $D = (x^1, y^1), ..., (x^N, y^N)$ consisting of input sentences x and output sentences y of length M or less.

$$L_{\mathrm{MLE}} = -\sum_{i=1}^{N}\sum_{t=1}^{M} \log p(y_t^i|y_{<t}^i, x^i) \qquad (2)$$

2.2 Fine-tuning

REINFORCE (Williams, 1992) is used for fine-tuning machine translation models based on reinforcement learning. REINFORCE is a type of policy gradient algorithm in which a machine translation model is trained to maximize the expected reward.

The loss function for fine-tuning is obtained by weighting the log-likelihood by the reward.

$$L_R = \sum_{i=1}^{N}\sum_{t=1}^{M} (R(\hat{y}^i) - R_b) \log p(\hat{y}_t^i|\hat{y}_{<t}^i, x^i), \qquad (3)$$

where h_t is the hidden state of the decoder at time t, R is the reward function, R_b is the baseline reward, and $\hat{y}^i$ is the output sentence from the decoder. In this study, the average reward within a mini-batch is used as the baseline reward.

To stabilize the training, the following loss function is used during reinforcement learning as in previous studies (Hashimoto and Tsuruoka, 2019).

$$L = \lambda L_{\mathrm{MLE}} + (1 - \lambda) L_R \qquad (4)$$

2.3 Rewards for Reinforcement Learning

In this study, the following evaluation metrics are used as rewards for reinforcement learning.

- BLEU[1] (Papineni et al., 2002) evaluates the surface token similarity between the output and reference sentences, using the word n-gram agreement rate.

[1] https://github.com/mjpost/sacrebleu

Reward	BLEU	Sent. BERT	BERT Reg.	SimCSE	chrF	BERTScore	BLEURT	STS BERT	Mean rank
None	33.73	75.66	0.0478	82.10	54.27	58.47	0.0639	3.654	7.75
BLEU	**34.26**	74.91	0.0202	81.93	**54.39**	58.01	0.0234	3.641	7.50
Sent. BERT	**33.78**	**75.79**	**0.0513**	**82.24**	**54.38**	**58.72**	**0.0649**	**3.656**	6.00
BERT Reg.	33.47	**75.80**	**0.0557**	**82.32**	54.25	**58.64**	**0.0681**	3.650	5.75
SimCSE	33.73	**75.84**	**0.0512**	**82.25**	**54.37**	**58.76**	**0.0669**	**3.659**	5.13
chrF	**33.90**	**75.81**	**0.0517**	**82.24**	**54.45**	**58.69**	**0.0671**	**3.657**	4.63
BERTScore	**33.96**	**75.80**	**0.0511**	**82.30**	**54.48**	**58.80**	**0.0677**	**3.658**	4.00
BLEURT	**33.85**	**75.90**	**0.0572**	**82.33**	**54.44**	**58.92**	**0.0759**	**3.660**	2.38
STS BERT	**34.09**	**76.11**	**0.0528**	**82.52**	**54.62**	**59.10**	**0.0700**	**3.684**	1.50

Table 1: Reinforcement learning performance of machine translation on IWSLT-2014 De→En task (bold indicates improvement by reinforcement learning, underlined indicates the highest value)

- chrF[1] (Popović, 2017) evaluates the surface token similarity between the output and reference sentences, using F1 scores of character n-grams and word n-grams.

- BERTScore[2] (Zhang et al., 2020) evaluates the semantic similarity between the output and reference sentences, using maximum matching of contextualized token embeddings obtained from pre-trained RoBERTa (`roberta-large`) (Liu et al., 2019).

- STS BERT (Yasui et al., 2019) evaluates the semantic similarity between the output and reference sentences, using BERT (Devlin et al., 2019) fine-tuned on the STS task (Cer et al., 2017).

- Sentence BERT[3] (Reimers and Gurevych, 2019) evaluates the semantic similarity between the output and reference sentences, using BERT fine-tuned on Natural Language Inference (NLI) task (Bowman et al., 2015).

- SimCSE[4] (Gao et al., 2021) evaluates the semantic similarity between the output and reference sentences, using RoBERTa fine-tuned by contrastive learning on sentence pairs with entailment labels in the NLI corpus as positive examples.

- BERT Regressor (Shimanaka et al., 2019) evaluates the semantic similarity between the output and reference sentences, using BERT fine-tuned on the metric task (Bojar et al., 2017).

- BLEURT[5] (Sellam et al., 2020) evaluates the semantic similarity between the output and reference sentences, using BERT pre-trained on an augmented data generated automatically by round-trip translation, and then fine-tuned on the metric task (Bojar et al., 2017).

3 Evaluation Experiments

3.1 Settings

IWSLT-2014 German-to-English task (Cettolo et al., 2014) was used for both pre-training and fine-tuning by reinforcement learning. The training dataset consists of $159,392$ sentence pairs, the validation dataset consists of $7,245$ sentence pairs, and the test dataset consists of $6,750$ sentence pairs.

Transformer (Vaswani et al., 2017) was used as the machine translation model, with 6 layers, 4 heads, 256 dimensions, and dropout rate of 0.3. In the pre-training, the optimization method was Adam (Kingma and Ba, 2015) (learning rate of 0.0003), the batch size was set to $2,048$, and the training was stopped by early stopping for BLEU on the validation data. In reinforcement learning, the optimization method was Adam (learning rate of 0.00001), $\lambda = 0.3$, batch size was 512, and training was stopped by early stopping for the evaluation metrics used as the reward. Reinforce-Joey[6] (Kiegeland and Kreutzer, 2021) was used for implementation.

The evaluation metrics in Section 2.3 were used for the reward calculation and the performance evaluation. STS BERT (Yasui et al., 2019) and BERT Regressor (Shimanaka et al., 2019) were im-

[2] `https://github.com/Tiiiger/bert_score`

[3] `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`

[4] `https://huggingface.co/princeton-nlp/sup-simcse-roberta-large`

[5] `https://storage.googleapis.com/bleurt-oss/bleurt-large-512.zip`

[6] `https://github.com/samuki/reinforce-joey`

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Mean
BLEU	0.412	0.413	0.565	0.393	0.460	0.531	0.524	0.471
chrF	0.517	0.531	0.671	0.525	0.599	0.607	0.591	0.577
STS BERT	0.535	0.597	0.667	0.637	0.611	0.589	0.608	0.606
Sentence BERT	0.632	0.621	0.692	0.685	0.690	0.657	0.635	0.659
SimCSE	0.696	0.628	0.684	0.696	0.713	0.660	0.672	0.678
BERTScore	0.710	0.745	0.833	0.756	0.746	0.751	0.775	0.759
BERT Regressor	0.712	0.732	0.858	0.804	0.775	0.789	0.765	0.776
BLEURT	**0.845**	**0.845**	**0.870**	**0.865**	**0.861**	**0.846**	**0.860**	**0.856**

Table 2: Pearson correlations with human evaluation in the WMT-2017 Metrics task (bold indicates the best score)

plemented using BERT$_{\text{BASE}}$[7] from HuggingFace Transformers[8] (Wolf et al., 2020).

3.2 Results

Table 1 shows the experimental results. The first line, “None”, is the baseline where only pre-training was performed without reinforcement learning. The comparison between the baseline and the reinforcement learning after the second line shows that the performance of all methods improved with reinforcement learning when the same evaluation metrics were used for both rewards and evaluation.

When BLEU was used as the reward, reinforcement learning improved only BLEU and chrF, i.e., surface-matching-based metrics, while performance deteriorated for the other BERT-based metrics. On the other hand, when chrF, also based on surface matching, was used as the reward, all evaluation metrics were improved by reinforcement learning.

Among the BERT-based rewards, reinforcement learning with Sentence BERT shows small improvement from the baseline model across the board, indicating that Sentence BERT is less effective. Reinforcement learning with SimCSE as the reward did not improve BLEU, and reinforcement learning with BERT Regressor as the reward resulted in worse BLEU than the baseline model.

Among the BERT-based rewards, we confirmed that the use of BERTScore, BLEURT, and STS BERT improved the performance of all the evaluation metrics tested in this study. In particular, STS BERT achieved the best performance on the majority of the evaluation metrics and was the most suitable reward function for reinforcement learning of machine translation.

4 Analysis

4.1 Meta-Evaluation of Evaluation Metrics

In this section, we examine whether the evaluation metrics that were effective as rewards for reinforcement learning in the experiments in Table 1 are highly correlated with the human evaluation of machine translation. In this analysis, we investigate the Pearson correlations between evaluation metrics and human evaluation for to-English language pairs in the WMT-2017 metrics task (Bojar et al., 2017). This task covers 7 language pairs: cs-en, de-en, fi-en, lv-en, ru-en, tr-en, and zh-en. Each 560 sentence pair (output and reference sentence pairs) is evaluated by human experts.

The results of the analysis are shown in Table 2. It can be seen that BERT-based evaluation metrics have a higher correlation with human evaluation than surface-matching metrics, BLEU and chrF. In particular, BLEURT shows the best correlation with human evaluation for all language pairs. However, contrary to expectations, STS BERT, which was the best reward for reinforcement learning, had a low correlation with human evaluation.

4.2 Correlations among Evaluation Metrics

In this section, we examine whether the correlations among the evaluation metrics affect the performance evaluation of reinforcement learning. As in Section 4.1, this section investigates the Pearson correlations among the metrics for to-English language pairs in the WMT-2017 metrics task.

The results are shown in Table 3. First, it can be seen that the correlation between BLEU and the other metrics was low. Although the correlation of BLEU with chrF, based on word n-gram match-

[7] https://huggingface.co/bert-base-uncased

[8] https://github.com/huggingface/transformers

	BLEU	STS BERT	chrF	SimCSE	Sent. BERT	BERT Reg.	BLEURT	BERTScore	Mean
BLEU	-	0.449	0.788	0.417	0.428	0.517	0.496	0.641	0.534
STS BERT	0.449	-	0.671	0.772	0.788	0.648	0.665	0.636	0.661
chrF	0.788	0.671	-	0.616	0.635	0.608	0.613	0.715	0.664
SimCSE	0.417	0.772	0.616	-	0.856	0.653	0.717	0.664	0.671
Sent. BERT	0.428	0.788	0.635	0.856	-	0.674	0.712	0.662	0.679
BERT Reg.	0.517	0.648	0.608	0.653	0.674	-	0.866	0.798	0.681
BLEURT	0.496	0.665	0.613	0.717	0.712	0.866	-	0.805	0.696
BERTScore	0.641	0.636	0.715	0.664	0.662	0.798	0.805	-	0.703

Table 3: Pearson's correlation coefficient between evaluation metrics

ing, and BERTScore, based on token-embedding matching, was relatively high, the correlation with sentence embedding-based metrics was low. These results indicates that BLEU may not be suitable sentence-based global evaluation. These characteristics of BLEU might have had effects on the low performance of BLEU in Tables 1 and 2.

Table 3 also indicates that the high performance of STS BERT in many of metrics as shown in Table 1 was unlikely due to the effect of compatibility between metrics because STS BERT tended to have relatively low correlations with other metrics.

5 Conclusion

In this study, we investigated BERT-based evaluation metrics as rewards for reinforcement learning in machine translation. The evaluation metrics can be used for both reward calculation and performance evaluation of machine translation. In the experiments, we examined the evaluation metrics in the total combination of using it as a reward and using it as a performance evaluation.

Experimental results on German-to-English translation of IWSLT-2014 show that reinforcement learning using BERT fine-tuned on STS task as a reward (STS BERT) can improve performance on many of evaluation metrics. The correlation between STS BERT and other evaluation metrics was relatively low, and this indicates that the high performance of STS BERT was unlikely due to the effect of metric compatibility. However, STS BERT has a relatively low correlation with human evaluation in the WMT-2017 metrics task and is not a good evaluation metric from this perspective.

BERTScore and BLEURT have high correlations with human evaluation and relatively high correlations with other evaluation metrics, and also improved all metrics as rewards for reinforcement learning. Therefore these metrics can also be considered good rewards.

As future work, we plan to use quality estimation (Specia et al., 2018) without reference sentences as a reward for reinforcement learning of machine translation. Rewards based on quality estimation have the potential to improve machine translation models in an unsupervised manner.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP20K19861. This research was also obtained from the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, pages 135–187.

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

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–17.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2019. Accelerated Reinforcement Learning for Sentence Generation by Vocabulary Prediction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3115–3125.

Samuel Kiegeland and Julia Kreutzer. 2021. Revisiting the Weaknesses of Reinforcement Learning for Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1673–1681.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Akifumi Nakamachi, Tomoyuki Kajiwara, and Yuki Arase. 2020. Text Simplification with Reinforcement Learning Using Supervised Rewards on Grammaticality, Meaning Preservation, and Simplicity. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 153–159.

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.

Maja Popović. 2017. chrF++: Words Helping Character N-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In *Proceedings of the 4th International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine Translation Evaluation with BERT Regressor. *arXiv:1907.12679*.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality Estimation for Machine Translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, pages 229–256.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-Sequence Learning as Beam-Search Optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin

Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A Study of Reinforcement Learning for Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621.

Go Yasui, Yoshimasa Tsuruoka, and Masaaki Nagata. 2019. Using Semantic Similarity as Reward for Reinforcement Learning in Sentence Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 400–406.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, pages 1–43.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.