An Empirical Study on Adversarial Attack on NMT: Languages and Positions Matter

Zhiyuan Zeng[1] and Deyi Xiong[2] *
School of New Media and Communication, Tianjin University, Tianjin, China[1]
College of Intelligence and Computing, Tianjin University, Tianjin, China[2]
{zhiyuan_zeng, dyxiong}@tju.edu.cn

Abstract

In this paper, we empirically investigate adversarial attack on NMT from two aspects: languages (the source vs. the target language) and positions (front vs. back). For autoregressive NMT models that generate target words from left to right, we observe that adversarial attack on the source language is more effective than on the target language, and that attacking front positions of target sentences or positions of source sentences aligned to the front positions of corresponding target sentences is more effective than attacking other positions. We further exploit the attention distribution of the victim model to attack source sentences at positions that have a strong association with front target words. Experiment results demonstrate that our attention-based adversarial attack is more effective than adversarial attacks by sampling positions randomly or according to gradients.

1 Introduction

Despite remarkable progress in recent years, neural machine translation (NMT) models are vulnerable to small perturbations (Cheng et al., 2018; Zhao et al., 2018). Adversarial training, which allows NMT models to learn from adversarial samples with perturbations, as a general approach, is widely used to improve the robustness of NMT (Ebrahimi et al., 2018; Vaibhav et al., 2019; Cheng et al., 2019, 2020a,a; Zou et al., 2019). Generally, NMT models yield target translations in an autoregressive way[1], which makes previous incorrectly predicted target tokens have a negative impact on future tokens to be generated. However, most approaches to generating NMT adversarial examples inject perturbations only into source sentences. Hence, are NMT models more vulnerable to adversarial attack on the source side? What roles do injecting perturbations into source sentences or into target translations play in improving the robustness of NMT?

The key interest of this paper is to attempt to answer these questions by an empirical and comparative study on different adversarial attacks on NMT models. First, we investigate adversarial attacks on the source side versus those on the target side. This study is to know which attack is more effective for NMT by measuring performance drop of the attacked models. Second, we empirically study the impact of attacking different positions on either source sentences or target translations to find whether NMT robustness is sensitive to positions. Third, based on the findings of the study, we propose a new adversarial attack generation method based on attention distribution.

Our contributions can be summarized as follows:

- By the study, we have empirically found that adversarial attack on the source side is more effective than that on the target side in terms of the performance degradation of NMT models under attack.
- We have further empirically found that adversarial attacks on front positions are more effective than those on back positions on the target side due to the autoregressive translation nature. We have also found that adversarial attacks on positions of the source side which are aligned to front positions of the target side are more effective than attacks on other positions on the source side.
- We propose a new adversarial attack generation approach that samples positions for injecting perturbations according to the attention distribution. Experiment results demonstrate that attention-based position sampling is more effective than random sampling and gradient-

*Corresponding author

[1]We leave the study of adversarial attack to non-autoregressive NMT models to our future work.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 454–460
August 1–6, 2021.

based sampling.

2 Related Work

Robustness is a well-known problem for neural networks (Szegedy et al., 2014; Goodfellow et al., 2015). Recent years have witnessed that many adversarial training approaches have been proposed to improve the robustness of NMT models. Cheng et al. (2018) generate adversarial samples at the lexical and feature level, and apply the adversarial learning to make adversarial samples natural. Zhao et al. (2018) utilize generative adversarial networks to generate adversarial examples that lie on the data manifold by searching in the semantic space of dense and continuous data representations. Ebrahimi et al. (2018) propose an attack framework for character-level NMT, which uses gradient to rank adversarial manipulations and to search for adversarial examples via either greedy search or beam search methods. Belinkov and Bisk (2018) attack character-level NMT by randomizing the order of letters or randomly replacing letters with their adjacent letters on the keyboard. Vaibhav et al. (2019) use back translation to generate adversarial samples that emulate natural noises. Cheng et al. (2020a) exploit the projected gradient method combined with gradient regularization to generate adversarial samples. Zou et al. (2019) employ reinforcement learning to decide which positions to attack. Tan et al. (2020) present a method to change inflectional morphology of words to craft plausible and semantically similar adversarial examples. Emelin et al. (2020) propose to generate adversarial examples by eliciting disambiguation errors.

All these approaches attack the source side of NMT in different ways. However distortions exist in not only the source language, but also the target language. This inspires us to compare the effectiveness of adversarial attack on the source and target side to NMT models. We have found that the NMT models are vulnerable to both the source and target attack. However, to our best knowledge, only Cheng et al. (2019) and Cheng et al. (2020b) take noises in target sentences into account. They generate adversarial samples for both source and target sentences. Their target-side adversarial examples are generated according to the attacked positions in corresponding source sentences, while their source-side adversarial samples are generated by randomly sampling positions to attack. We improve their method by attacking the source side according to the attention distribution. Experiments validate the effectiveness of our method.

3 Data and Setup

We conducted experiments on two translation tasks: English-Chinese and English-Japanese. Data for English-Chinese translation are from the United Nations English-Chinese corpus (Ziemski et al., 2016). We built the training/validate/test set for this task by randomly sampling 3M/2K/2K sentence pairs from the whole corpus. For the English-Japanese translation task, we aggregated the training set of KFTT (Neubig, 2011), JESC (Pryzant et al., 2018) and TED talks (Cettolo et al., 2012) as our training set, which consists of 3.9M sentence pairs. We evaluated our models on the validation set and test set of KFTT (Neubig, 2011). We split words into sub-word units with subword regularization (Kudo, 2018) and built a shared vocabulary of 32K subwords for both English-Chinese and English-Japanese.

We used the base Transformer model (Vaswani et al., 2017) with 512 hidden units as the victim model. The hyper-parameters of the base Transformer follows the default setting in Vaswani et al. (2017). We implemented the adversarial attack and training methods of Cheng et al. (2019) and followed their hyper-parameter setting in our experiments. The details of our implementation is shown in Section 4.

We injected perturbations into either source sentences or target sentences to generate adversarial examples which were used to evaluate NMT models. Since we could not inject perturbations into the target inputs of NMT models at the test time, we evaluated NMT models with target-side adversarial samples at training time on the validation dataset. Except where otherwise specified, the performance of the victim model was measured by word accuracy on the validation data. If we evaluated the victim model on the test set, detokenized BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020) were reported. Although the target-side inputs of NMT models could not be attacked at test time, there still exists noise or errors in them due to error propagation in the autoregressive decoding. Evaluating NMT with perturbated target sentences at training time enables us to analyze the vulnerability of NMT to the noise in target-side inputs, and inspires us to improve the robustness of NMT models to such noise.

src-tgt	en-zh	zh-en	en-ja	ja-en
noisy-clean	**55.79**	**61.89**	**50.21**	**51.98**
clean-noisy	61.32	64.00	52.74	52.58
clean-clean	71.16	78.76	60.35	62.30
noisy-noisy	46.55	46.55	40.63	40.78

Table 1: Word Translation accuracy of victim model under the adversarial attack on the source (src) vs. target (tgt)

4 Implementation Details

The adversarial attack and training framework used in this paper is based on Cheng et al. (2019). They inject perturbations into the source/target sentences by replacing a word in a sentence with the words that are semantically similar to the words being replaced. Words to be replaced in a source sentence are sampled according to the uniform distribution, while those in a target sentence are sampled according to the attention distribution. We tried three different ways to sample words to inject perturbations into source sentence in Section 8. Given a word to be replaced, Cheng et al. (2019) use a bi-directional language model to choose candidate words from vocabulary which share similar semantics to it, and then use gradients to search a word from candidate words to replace it. Cheng et al. (2019) combine a left-to-right and right-to-left language model to rank candidate words, while we combine the two uni-directional language models by multiplying their likelihood for simplicity. Cheng et al. (2019) train their NMT models with both clean data and adversarial samples from scratch. To save training time, we pretrain our NMT models with clean data before adversarial training.

5 Adversarial Attack on Source vs. Target

In this section, we compare the effect of the source and target attack according to the performance of the victim model. We adversarially inject perturbations into source sentences and keep target translations unchanged (clean) for the source attack while the target attack works the other way around. Our adversarial examples for both the source and target attack are generated by the method of Cheng et al. (2019). To make the comparison between the source and target attack fair, we randomly sample positions to attack for both of them.

Results are shown in Table 1. The NMT model with noisy source and clean target performs worse

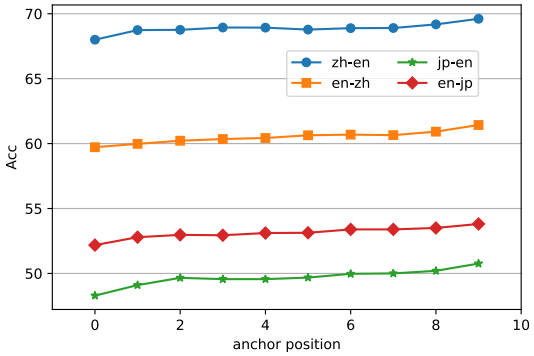

Figure 1: The word translation accuracy of an NMT model under attack at different anchor positions on the target side.

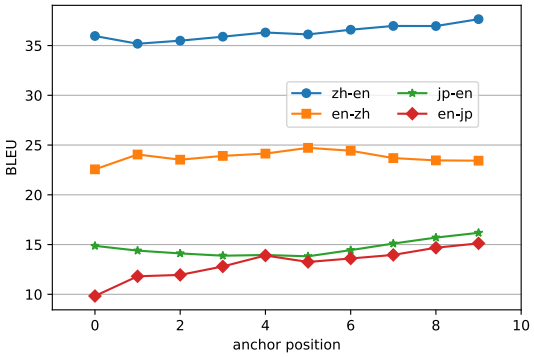

Figure 2: The BLEU score of an NMT model under attack at different anchor positions on the source side.

than that with clean source and noisy target on all translation tasks, in terms of word translation accuracy, which indicates that the source attack is more effective than the target attack. We also oberve that the adversarial attack on the source together with target side is much better than that on a single side, therefore we suggest that adversarial attacks on both the source and target side should be conducted to deploy a robust NMT system.

6 Adversarial Attack at Different Positions

In this section, we investigate the impact of attacked positions in the source and target sentences on NMT. We start with adversarial attack on the target side. Adversarial attacks at the front of a target sentence are supposed to be more effective than those at the end of the target sentence, since noises in the front of the target sentence will negatively affect future target tokens, while noises at the end of the target sentence could not affect already

generated tokens for a left-to-right decoder.

Given a sentence of length L, we uniformly select 10 anchor positions from the sentence:

$$\hat{x_j} = [\frac{L \times j}{10}] \quad (1)$$

where $\hat{x_j}$ is the jth anchor position ($0 <= j < 10$), $[\cdot]$ is a rounding operation. For each anchor position $\hat{x_j}$, we sample several positions close to it according to the discrete Gaussian distribution, which is formulated as:

$$p(x) = \frac{e^{-(x-\hat{x})^2}}{\sum_{i=0}^{L-1} e^{-(i-\hat{x})^2}} \quad (2)$$

where $p(x)$ is the probability that position x is attacked, $\hat{x}$ is the anchor position that we want the sampled positions to surround. The denominator normalizes the sum of the probabilities to 1.

Results of adversarial attack on different anchor positions on the target side are shown in Figure 1. On all translation tasks, the word translation accuracy of the victim model goes up as attacked positions move from the starting position to the end of target sentences, which confirms that attacking at the front of a target sentence is more effective than attacking at the end.

We also perform adversarial attack on source sentences at different anchor positions. Results are displayed in Figure 2. We measure the performance of the victim model for the source attack at different positions on the test set with the metric of BLEU. On both English-Chinese and English-Japanese tasks, BLEU scores go up as the attacked positions move from the start to the end of source sentences, which indicates that attacking the front of a source sentence is also more effective than attacking the end for both English-Chinese and English-Japanese translation. We suppose that the reason for this is that words at front positions of source sentences usually align to words at front positions of target sentences for the two language pairs. Experiment results in Section 7 empirically validate this hypothesis.

7 Attention Weights at Different Positions

In section 6, we suppose that words at front positions of source sentences usually align to words at front positions of target sentences for both English-Chinese and English-Japanese. We empirically validate this by comparing the attention weights from

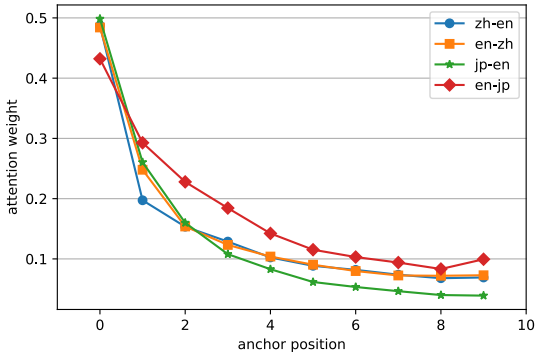

Figure 3: Attention weights from different anchor positions of source sentences to the first token of target sentences.

source tokens at different positions to the first target token. Following the sampling technique in section 6, we uniformly select 10 anchor positions from a source sentence and then sample positions surrounding these anchor positions according to the distribution formulated in Eq (2). For every anchor position, we report the sum of attention weights from the sampled source tokens around the anchor position to the first target token. The results are shown in Figure 3. As expected, the attention weights from the sampled source tokens to the first target token go down as the their corresponding positions move from the start to the end of the source sentence in both English-Chinese and English-Japanese translation, which confirms that words at front positions of source sentences have a stronger association with words at front positions of target sentences than other positions for the two language pairs.

8 Adversarial Attack based on Attention Distribution

In Section 6, we have found that generating perturbations at front positions on the target side is more effective than attacking other positions. As attention weights in NMT models can be seen as the strength of association between the source and target tokens (Bahdanau et al., 2015). Hence we sample positions of a source sentence to inject perturbations according to the attention distribution. Particularly, the query used to produce the attention distribution is the representation of the first target token and the key is the set of representations of source tokens. There are multiple cross-attention heads in Transformer, each of which produces an

attack / model	rand		grad		attn	
	BELU	BERTScore	BELU	BERTScore	BELU	BERTScore
victim	19.2	83.8	22.2	84.7	**17.4**	**83.4**
train-rand	25.0	86.1	28.9	87.1	**22.2**	**85.6**
train-grad	23.6	85.7	28.6	87.1	**21.3**	**85.2**
train-attn	24.0	85.7	27.7	86.8	**23.3**	85.8

Table 2: BLEU and BERTScore of the victim model and three adversarially trained models. "rand", "grad" and "attn" indicates that adversarial examples are generated at attacked positions sampled randomly, according to gradients and attention distribution, respectively. "train-X" denotes that NMT models are adversarially trained with adversarial examples generated by the "X" method. The models were evaluated on the test set of the English-Chinese corpus.

attention matrix. The average of attention distributions of all heads is hence used for attacking.

We compare our proposed attention-based attack with attacks that either randomly sample source positions or sample positions according to gradients. For gradient-based sampling, we follow Liang et al. (2018) to estimate the L_∞ norm of the gradient of a word embedding as the importance score of the corresponding word, and then sample positions to attack from the normalized importance score. We have implemented the three adversarial attack methods based on the framework proposed in Cheng et al. (2019).[2] The only difference of these methods is that they use different ways to sample positions to attack. We also use the adversarial training method proposed in Cheng et al. (2019) to fine-tune NMT model with adversarial samples generated with the three attacking methods.

BLEU scores and BERTScores of the three adversarially trained models on the test set are shown in Table 2. It can be seen that BLEU scores and BERTScores of almost all models under our proposed attack ("attn") are lower than those under the other two attacking methods, which indicates the superiority of the proposed attention-based attack over the other two attack methods. It is surprising that the attack that samples positions according to the gradient ("grad") is not better than the attack that samples from a uniform distribution ("rand"), which may suggest that the L_∞ norm of the gradient cannot measure the importance of a word in a sentence. We can further extend our method to the black-box attack with the alignment from SMT models (Och and Ney, 2003), which is left to our future work. Our attention-based attack is proposed for autoregressive NMT models that generate target translations from left to right. It will not work for non-autoregressive NMT models (Gu et al., 2017) or autoregressive NMT models that generates translations in an arbitrary order (Stern et al., 2019).

[2]Cheng et al. (2019) randomly sample positions to attack source sentences in their paper.

9 Conclusion

In this paper, we have empirically investigated adversarial attack on NMT models. We compare adversarial attack on the source vs. target side, and find that the former is more effective than the latter. We also study adversarial attack at different positions in either source or target sentences, and observe that attacking front positions in either source or target sentences for English-Chinese and English-Japanese translation is more effective than attacking back positions. We further exploit attention distribution to attack words of a source sentence at positions that have a high association with words at front positions of the corresponding target sentence. Experiments validate the effectiveness of our proposed attention-based attack.

Acknowledgements

The present research was partially supported by OPPO. We would like to thank the anonymous reviewers for their insightful comments.

References

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*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC,*

Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation, EAMT 2012, Trento, Italy, May 28-30, 2012*, pages 261–268. European Association for Machine Translation.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020a. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3601–3608. AAAI Press.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4324–4333. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020b. Advaug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5961–5970. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1756–1766. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 653–663. Association for Computational Linguistics.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7635–7653. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *CoRR*, abs/1711.02281.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4208–4215. ijcai.org.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguistics*, 29(1):19–51.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: japanese-english subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5976–5985. PMLR.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Samson Tan, Shafiq R. Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 2920–2935. Association for Computational Linguistics.

Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1916–1920. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2019. A reinforced generation of adversarial samples for neural machine translation. *CoRR*, abs/1911.03677.