Generating Authentic Adversarial Examples beyond Meaning-preserving with Doubly Round-trip Translation

Anonymous ACL submission

Abstract

Generating adversarial examples for Neural Machine Translation (NMT) with single Round-Trip Translation (RTT) has achieved promising results by releasing the meaning-preserving restriction. However, a potential pitfall for this approach is that we cannot decide whether the generated examples are adversarial to the target NMT model or the auxiliary backward one, as the reconstruction error through the RTT can be related to either. To remedy this problem, we propose a new definition for NMT adversarial examples based on the Doubly Round-Trip Translation (DRTT). Specifically, apart from the source-target-source RTT, we also consider the target-source-target one, which is utilized to pick out the authentic adversarial examples for the target NMT model. Additionally, to enhance the robustness of the NMT model, we introduce the masked language models to construct bilingual adversarial pairs based on DRTT, which are used to train the NMT model directly. Extensive experiments on both the clean and noise test sets (including the artificial and natural noise) show that our approach substantially improves the robustness of NMT models.[1]

1 Introduction

In recent years, neural machine translation (NMT) (Cho et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017) has achieved rapid advancement in the translation performance (Yang et al., 2020; Lu et al., 2021). However, the NMT model is not always stable enough, as its performance can drop significantly when small perturbations are added into the input sentences (Belinkov and Bisk, 2017; Cheng et al., 2020). Such perturbed inputs are often referred to as adversarial examples in the literature, and how to effectively generate and utilize adversarial examples for NMT is still an open question.

[1]Codes will be publicly available once accepted.

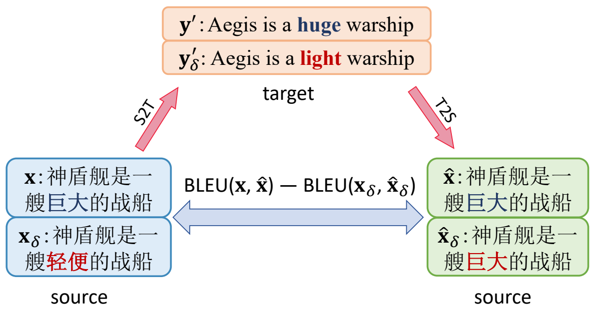

Figure 1: An example of the source-target-source RTT process on a perturbed input $\mathbf{x}_\delta$ by replacing “巨大 (huge)” to “轻便 (light)”.

Conventional approaches (Ebrahimi et al., 2018; Cheng et al., 2019) for generating NMT adversarial examples always follow the meaning-preserving assumption, i.e., an NMT adversarial example should preserve the meaning of the source sentence but destroy the translation performance drastically (Michel et al., 2019; Niu et al., 2020). With the meaning-preserving restriction, the researchers try to add perturbations on the source inputs as small as possible to ensure the meaning of the source sentence is unchanged, which severely limits the search space of the adversarial examples. Additionally, it is much problematic to craft a minor perturbation on discrete text data, since some random transformations (e.g., swap, deletion and replacement) may change, or even reverse semantics of the text data, breaking the aforementioned meaning-preserving assumption. To break this limitation, Zhang et al. (2021) introduce a new definition for NMT adversarial examples: *an effective NMT adversarial example imposes minor shifting on the source and degrades the translation dramatically, would naturally lead to a semantic-destroyed round-trip translation result*. Take the case in Figure 1 as an example: $\mathbf{x}_\delta$ reverses the semantics of input $\mathbf{x}$ by replacing “巨大 (huge)” to “轻便 (light)”. Since the semantics of $\mathbf{x}$ and $\mathbf{x}_\delta$ are completely different, it is unreasonable to use the original target sentence of $\mathbf{x}$ to evaluate the attacks directly. Therefore, Zhang et al. (2021) propose to evaluate the BLEU score between $\mathbf{x}_\delta$ and its reconstructed sentence $\hat{\mathbf{x}}_\delta$ from the source-target-source round-trip translation (RTT), as well as the BLEU score between the original sentence $\mathbf{x}$ and its reconstructed sentence $\hat{\mathbf{x}}$. They take the decrease between the two BLEU scores mentioned above as the adversarial effect. Specifically, if the BLEU decrease exceeds a predefined threshold, $\mathbf{x}_\delta$ is concluded to be an adversarial example for the target NMT model.

While achieving promising results by breaking the meaning-preserving limitation, there are two potential pitfalls in the work of Zhang et al. (2021): (1) Since the source-target-source RTT involves two stages, i.e., the source-to-target translation (S2T) performed by the target NMT model and target-to-source translation (T2S) performed by an auxiliary backward NMT model, we cannot decide whether the BLEU decrease is really caused by the target NMT model. As we can see from the example in Figure 1, the translation from $\mathbf{x}_\delta$ to $\mathbf{y}'_\delta$ is pretty good, but the translation from $\mathbf{y}'_\delta$ to $\hat{\mathbf{x}}_\delta$ is really poor. We can conclude that the BLEU decrease is actually caused by the auxiliary backward model and thus $\mathbf{x}_\delta$ is not the adversarial example for the target NMT model. Even if Zhang et al. (2021) try to mitigate this problem by fine-tuning the auxiliary backward model on the test sets, we find this problem still remains. (2) They only generate the monolingual adversarial examples on the source side to attack the NMT model, without proposing methods on how to defend these adversaries and improve the robustness of the NMT model.

To address the issues mentioned above, we first propose a new definition for NMT adversarial examples based on Doubly Round-Trip Translation (DRTT), which can ensure the examples that meet our definition are the authentic adversarial examples for the target NMT model. Specifically, apart from the source-target-source RTT (Zhang et al., 2021), we additionally consider a target-source-target RTT on the target side. The main intuition is that an effective adversarial example for the target NMT model shall cause a large BLEU decrease on the source-target-source RTT while maintaining a small BLEU decrease on target-source-target RTT. Based on this definition, we craft the candidate adversarial examples with the source-target-source RTT as Zhang et al. (2021), and then pick out the authentic adversaries with the target-source-target RTT. Furthermore, to solve the second problem, we introduce the masked language models (MLMs) to construct the bilingual adversarial pairs by performing phrasal replacement on the generated monolingual adversarial examples and the original target sentences synchronously, which are then utilized to train the NMT model directly. Experiments on both clean and noise test sets (including five types of artificial and nature noise) show that the proposed approach not only generates effective adversarial examples, but also improves the robustness of the NMT model over all kinds of noises. To conclude, our main contributions are summarized as follows:

- We propose a new definition for NMT adversarial examples based on the doubly round-trip translation, which can pick out the authentic adversarial examples for the target NMT model.

- We introduce the masked language models to construct the bilingual adversarial pairs, which are then utilized to improve the robustness of the NMT model.

- Extensive experiments show that the proposed approach not only improves the robustness of the NMT model on both artificial and natural noise, but also performs well on the clean test sets.

2 Related Work

2.1 Adversarial Examples for NMT

The previous approaches for constructing NMT adversarial examples can be divided into two branches: white-box and black-box. The white-box approaches are based on the assumption that the architecture and parameters of the NMT model are accessible (Ebrahimi et al., 2018; Cheng et al., 2019; Chen et al., 2021). These methods usually achieve superior performance since they can construct and defend the adversaries tailored for the target NMT model. However, in the real application scenario, it is always impossible for us to access the inner architecture of the model. On the contrary, the black-box approaches never access to inner architecture and parameters of the model. In this line, Belinkov and Bisk (2017) rely on synthetic and naturally occurring language error to generate adversarial examples. Recently, Zhang et al. (2021) craft adversarial examples beyond the meaning-preserving restriction with the round-trip

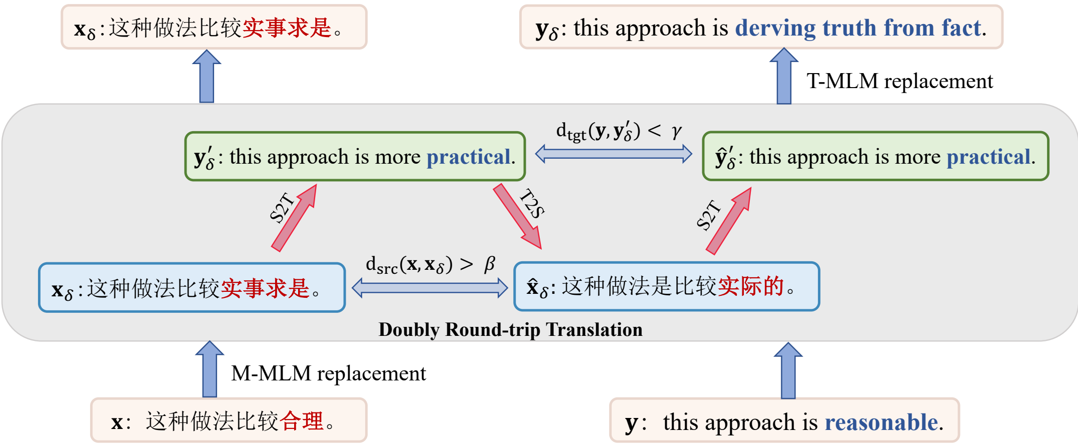

Figure 2: The overview of the bilingual adversarial pair generation under the definition of DRTT. $(\mathbf{x}, \mathbf{y})$ denote the source and target sentence. $(\mathbf{x}_\delta, \mathbf{y}_\delta)$ denote the generated bilingual adversarial pair.

translation. Our work builds on top of Zhang et al. (2021), and it applies the doubly round-trip translation to generate the authentic adversarial examples.

2.2 Masked Language Model

Masked Language Model (MLM) (Devlin et al., 2018; Conneau and Lample, 2019) has achieved state-of-the-art results on many monolingual and cross-lingual language understanding tasks. MLM randomly masks some of the tokens in the input, and then predicts those masked tokens. Recently, some work adopt MLM to do word replacement as a data augmentation strategy. Jiao et al. (2019) leverage an encoder-based MLM to predict word replacements for single-piece words. Liu et al. (2021) construct augmented sentence pairs by sampling new source phrases and corresponding target phrases with transformer-based MLMs. Following Liu et al. (2021), we introduce the transformer-based MLMs to construct the bilingual adversarial pairs. The main difference between our work and Liu et al. (2021) is that we choose to mask the adversarial phrases or words at each step and Liu et al. (2021) mask the words randomly.

3 Method

In this section, we first describe our proposed definition for NMT adversarial examples, and then present the way of constructing the bilingual adversarial pairs.

3.1 Adversarial Examples for NMT

For clarity, we first introduce the traditional definitions for NMT adversarial examples, i.e., the definitions based on the meaning-preserving and RTT, and then elaborate our new definition based on DRTT. We will use the following notations: $\mathcal{X}$ and $\mathcal{Y}$ refer to the source and target training space, respectively. $M_{x\to y}$ is the target NMT model trained from source-to-target mapping $f\colon \mathcal{X} \to \mathcal{Y}$, while $M_{y\to x}$ is an auxiliary target-to-source model trained from mapping $g\colon \mathcal{Y} \to \mathcal{X}$. $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ denotes the source and target sentence, respectively. $\mathbf{x}_\delta$ and $\mathbf{y}_\delta$ denote the perturbed version of $\mathbf{x}$ and $\mathbf{y}$, respectively. $\mathbf{y}' = f(\mathbf{x})$ and $\mathbf{y}'_\delta = f(\mathbf{x}_\delta)$ are the forward translations generated by $M_{x\to y}$. $\hat{\mathbf{x}} = g(f(\mathbf{x}))$ and $\hat{\mathbf{x}}_\delta = g(f(\mathbf{x}_\delta))$ are reconstructed sentences generated with the source-target-source RTT. $\hat{\mathbf{y}} = f(g(\mathbf{y}))$ and $\hat{\mathbf{y}}'_\delta = f(g(\mathbf{y}'_\delta))$ are reconstructed sentences generated with the target-source-target RTT. $\mathrm{sim}(\cdot,\cdot)$ is a function for evaluating the similarity of two sentences, and we use BLEU (Papineni et al., 2002a) as the similarity function in this paper.

Definition based on meaning-preserving. Suppose $M_{x\to y}$ translates the input $\mathbf{x}$ and its perturbed version $\mathbf{x}_\delta$ to $\mathbf{y}'$ and $\mathbf{y}'_\delta$, respectively. $\mathbf{x}_\delta$ is an adversarial examples when it meets:

$$\begin{cases} \mathrm{sim}(\mathbf{x}, \mathbf{x}_\delta) > \eta, \\ \mathrm{sim}(\mathbf{y}, \mathbf{y}') - \mathrm{sim}(\mathbf{y}, \mathbf{y}'_\delta) > \alpha, \end{cases} \quad (1)$$

where η is a threshold to ensure a high similarity between $\mathbf{x}_\delta$ and $\mathbf{x}$, so that they can meet the meaning-preserving restriction. A larger α indicates a more strict definition of the NMT adversarial example.

Definition based on RTT. Zhang et al. (2021) point out that the perturbation δ may change, even reverse the meaning of $\mathbf{x}$ to some extent, so it is incorrect to use $\mathbf{y}$ as a target sentence to measure the semantic destruction on the target side. Therefore, they introduce the definition based on RTT which gets rid of the meaning-preserving restriction. The percentage decrease of similarity between $\mathbf{x}$ and $\mathbf{x}_\delta$ through the source-target-source RTT is regarded as the adversarial effect $\mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta)$, is calculated as:

$$\mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta) = \frac{\mathrm{sim}(\mathbf{x}, \hat{\mathbf{x}}) - \mathrm{sim}(\mathbf{x}_\delta, \hat{\mathbf{x}}_\delta)}{\mathrm{sim}(\mathbf{x}, \hat{\mathbf{x}})}. \quad (2)$$

A large $\mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta)$ indicates that the perturbed sentence $\mathbf{x}_\delta$ can not be well reconstructed by RTT when compared to the reconstruction quality of the original source sentence $\mathbf{x}$, so $\mathbf{x}_\delta$ is likely to be an adversarial example.

Definition based on DRTT. In Eq.(2), $\mathrm{sim}(\mathbf{x}, \hat{\mathbf{x}})$ is a constant value given the input $\mathbf{x}$ and the NMT models. Therefore, the $\mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta)$ is actually determined by $-\mathrm{sim}(\mathbf{x}_\delta, \hat{\mathbf{x}}_\delta)$, which can be interpreted as the reconstruction error between $\mathbf{x}_\delta$ and $\hat{\mathbf{x}}_\delta$. As we all known, the reconstruction error can be caused by two independent translation processes: the forward translation process f performed by the target NMT model and the backward translation process g performed by the auxiliary backward model. Therefore, there may be three occasions when we get a large $\mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta)$: 1) A large semantic destruction in $f(\mathbf{x}_\delta)$ and a small semantic destruction in $g(\mathbf{y}'_\delta)$; 2) A large semantic destruction in $f(\mathbf{x}_\delta)$ and a large destruction in $g(\mathbf{y}'_\delta)$; 3) A small semantic destruction in $f(\mathbf{x}_\delta)$ and a large destruction in $g(\mathbf{y}'_\delta)$. We can conclude $\mathbf{x}_\delta$ is an adversarial example for the target NMT model in occasion 1 and 2, but not in occasion 3. Therefore, the definition based on RTT may contain many fake adversarial examples.

To address this problem, we add a target-source-target RTT starting from the target side. The percentage decrease of the similarity between $\mathbf{y}$ and $\mathbf{y}'_\delta$ through the target-source-target RTT, denoted as $\mathrm{d}_{\mathrm{tgt}}(\mathbf{y}, \mathbf{y}'_\delta)$, is calculated as:

$$\mathrm{d}_{\mathrm{tgt}}(\mathbf{y}, \mathbf{y}'_\delta) = \frac{\mathrm{sim}(\mathbf{y}, \hat{\mathbf{y}}) - \mathrm{sim}(\mathbf{y}'_\delta, \hat{\mathbf{y}}'_\delta)}{\mathrm{sim}(\mathbf{y}, \hat{\mathbf{y}})}. \quad (3)$$

We take both $\mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta)$ and $\mathrm{d}_{\mathrm{tgt}}(\mathbf{y}, \mathbf{y}'_\delta)$ into consideration and define $\mathbf{x}_\delta$ as an adversarial examples when it meets:

$$\begin{cases} \mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta) > \beta, \\ \mathrm{d}_{\mathrm{tgt}}(\mathbf{y}, \mathbf{y}'_\delta) < \gamma. \end{cases} \quad (4)$$

The interpretation is intuitive: if $\mathrm{d}_{\mathrm{tgt}}(\mathbf{y}, \mathbf{y}'_\delta)$ is lower than the threshold γ, we can conclude that the reconstruction error between $\mathbf{y}'_\delta$ and $\hat{\mathbf{y}}'_\delta$ is very low. Namely, we can ensure a small semantic destruction of $g(\mathbf{y}'_\delta)$. Therefore, if $\mathrm{d}_{\mathrm{src}}(\mathbf{x}, \mathbf{x}_\delta)$ is larger than β, we can conclude the BLEU decrease through the source-target-source RTT is caused by the target NMT model, so that we can conclude $\mathbf{x}_\delta$ is an authentic adversarial example.

3.2 Bilingual Adversarial Pair Generation

Since the proposed definition breaks the meaning-preserving restriction, the adversarial examples may be semantically distant from the original source sentence. Thus, we cannot directly pair the adversarial examples with the original target sentences. In this section, we propose our approach for generating bilingual adversarial pairs, which performs the following three steps: 1) Training Masked Language Models: using monolingual and parallel data to train masked language models; 2) Phrasal Alignment: obtaining alignment between the source and target phrases; 3) Phrasal Replacement: generating bilingual adversarial pairs by performing phrasal replacement on the source and target sentences synchronously with the trained masked language models. The whole procedure is illustrated in Figure 2.

Training Masked Language Models. We train two kinds of masked language models, namely monolingual masked language model (M-MLM) and translation masked language model (T-MLM), for phrasal replacement on the source and target sentence, respectively. The M-MLM introduces a special [MASK] token which randomly masks some of the tokens from the input in a certain probability, and the objective is to predict the original masked words. Following Liu et al. (2021), we train the M-MLM on monolingual datasets and use an encoder-decoder Transformer model (Vaswani et al., 2017) to tackle the undetermined number of tokens during generation. The T-MLM takes the identical model structure and similar training process as the M-MLM. The main difference is that T-MLM relies on the parallel corpus. T-MLM concatenates parallel sentences by a special token [SEP] and only masks words on the target side. The objective of T-MLM is to predict the original masked words on the target side.

Phrasal Alignment. Phrasal alignment projects each phrase in the source sentence $\mathbf{x}$ to its alignment phrase in the target sentence $\mathbf{y}$. We first generate the alignment between $\mathbf{x}$ and $\mathbf{y}$ using FastAlign[2]. Then we extract the phrase-to-phrase alignment through running the phrase extraction algorithm of NLTK[3], and get a mapping function p.

Phrasal Replacement. Given the source sentence $\mathbf{x} = \{s_1, s_2, \dots, s_n\}$ and the target sentence $\mathbf{y} = \{t_1, t_2, \dots, t_m\}$, s_i is the i-th phrase in $\mathbf{x}$, $t_{p(i)}$ is the $p(i)$-th phrase in $\mathbf{y}$ which is aligned to s_i by the mapping function p. we construct the candidate bilingual adversarial pairs $(\mathbf{x}_\delta, \mathbf{y}_\delta)$ by performing the phrasal replacement on $(\mathbf{x}, \mathbf{y})$ repeatedly until c percentage phrases in $\mathbf{x}$ have been replaced.

Here, we take the replacing process for s_i and $t_{p(i)}$ as an example. Considering the not attacked yet phrase s_i in $\mathbf{x}$, we first build a candidate set $\mathcal{R}_i = \{r_i^1, r_i^2, \dots, r_i^k\}$ for s_i with the prepared M-MLM. Specifically, we extract the k candidate phrases with top k highest predicted probabilities by feeding $\mathbf{x}^{\setminus i}$ into M-MLM, where $\mathbf{x}^{\setminus i}$ is the masked version of $\mathbf{x}$ by masking s_i. We select the best candidate r_i^* for s_i as:

$$r_i^* = \underset{j \in \{1, \cdots, k\}}{\arg\max}\, \mathrm{d}_{\text{src}}(\mathbf{x}, \mathbf{x}^{\setminus i:j}), \tag{5}$$

where $\mathbf{x}^{\setminus i:j}$ is the noised version by replacing s_i with r_i^j. With s_i being replaced, we need to replace $t_{p(i)}$ to ensure they are still semantically aligned. To this end, we feed the concatenation of $\mathbf{x}^{\setminus i:*}$ and $\mathbf{y}^{\setminus p(i)}$ into T-MLM, and choose the output phrase with the highest predicted probability as the substitute phrase for $t_{p(i)}$.

Finally, to decide whether $(\mathbf{x}_\delta, \mathbf{y}_\delta)$ is an authentic bilingual adversarial pair for the target NMT model, we perform a target-source-target RTT starting from the target side and calculate $\mathrm{d}_{\text{tgt}}(\mathbf{y}, \mathbf{y}'_\delta)$ between $\mathbf{y}'_\delta$ and its reconstruction sentence $\hat{\mathbf{y}}'_\delta$ according to Eq.(4). We take $(\mathbf{x}_\delta, \mathbf{y}_\delta)$ as an authentic bilingual adversarial pair if $\mathrm{d}_{\text{src}}(\mathbf{x}, \mathbf{x}_\delta)$ is greater than β and $\mathrm{d}_{\text{tgt}}(\mathbf{y}, \mathbf{y}'_\delta)$ is less than γ. We formalize these steps in Algorithm 1 in Appendix A.

4 Experimental Settings

We evaluate our model under artificial noise in Zh→En and En→De translation tasks, and under natural noise in En→Fr translation task. The details of the experiments are elaborated in this section.

4.1 Dataset

For the Zh→En task, we use the LDC corpus with 1.25M sentence pairs for training[4], NIST06 for validation, and NIST 02, 03, 04, 05, 08 for testing. For the En→De task, we use the publicly available dataset WMT'17 En-De (5.85M) for training, and take the *newstest16* and *newstest17* for validation and testing, respectively. In En→Fr task, we follow Liu et al. (2021) to combine the WMT'19 En→Fr (36k) robustness dataset with Europarl-v7 (2M) En-Fr pairs for training. We take the development set of the MTNT (Michel and Neubig, 2018) for validation and the released test set of the WMT'19 robustness task for testing. As for MLMs, we use the Chinese sentences of the parallel corpus to train the Chinese M-MLM, and use the whole parallel corpus to train Zh-En T-MLM. We train the English M-MLM with News Commentary and News Crawl 2010 (7.26M in total) monolingual corpus following Liu et al. (2021). T-MLM for En-De and En-Fr are trained with their original parallel corpus.

4.2 Model Configuration and Pre-processing

The MLMs and NMT models in this paper take Transformer-base (Vaswani et al., 2017) as the backbone architecture. We implement all models base on the open-source toolkit Fairseq[5]. As for hyper-parameters, β is set to 0.01 and γ is set to 0.5 for Zh→En. For En→De and En→Fr, β and γ is set to 0.5 and 0.5, respectively. The replacement ratio c is set to 0.2 following Liu et al. (2021), and the candidate number k is set to 1. The details of model configuration and the number of the generated adversarial examples are shown in the Appendix B. Following previous work, the Zh→En performance is evaluated with the BLEU (Papineni et al., 2002b) score calculated by *multi-bleu.perl* script. For En→De and En→Fr, we use SacreBLEU (Post, 2018) for evaluation[6].

4.3 Comparison Methods

To test the effectiveness of our model, we take the following three systems as comparison methods:

Baseline: The vanilla Transformer model for NMT (Vaswani et al., 2017). In our work, we

[2] https://github.com/clab/fast_align

[3] https://github.com/nltk/nltk/blob/develop/nltk/translate/phrase_based.py

[4] It is extracted from LDC data, including LDC 2002E18, 2003E07, 2003E14, 2004T08 and 2005T06.

[5] https://github.com/pytorch/fairseq

[6] nrefs:1 | case:mixed | eff:no | tok:intl | smooth:exp | version:2.0.0

Noise	Model	Zh→En				En→De			
		0.1	0.2	0.3	AVG	0.1	0.2	0.3	AVG
Deletion	baseline	32.98	26.59	20.54	26.70	19.82	13.71	9.33	14.29
	+TCWR	34.47	27.76	21.38	27.87	19.61	13.77	9.08	14.15
	+RTT	33.84	27.43	20.74	27.33	19.61	13.48	9.27	14.12
	+DRTT(ours)	**35.10**∗∗	**28.12**∗	**22.07**∗∗	**28.43**	**19.83**	**14.22**	**9.48**	**14.51**
Swap	baseline	36.14	32.88	30.21	33.08	21.47	16.97	**13.21**	17.22
	+TCWR	37.67	34.15	31.47	34.43	20.52	16.31	12.80	16.54
	+RTT	37.14	34.34	31.42	34.30	20.23	15.47	11.52	15.74
	+DRTT(ours)	**37.90**∗	**34.65**	**31.92**∗	**34.82**	**21.51**∗∗	**17.36**∗∗	12.91∗∗	**17.26**
Insertion	baseline	39.96	39.10	38.41	39.16	26.86	**26.54**	25.48	25.96
	+TCWR	41.32	40.07	39.60	40.33	26.27	25.55	24.33	25.38
	+RTT	41.75	40.82	39.90	40.82	26.18	25.06	23.68	24.97
	+DRTT(ours)	**41.98**	**40.90**	**40.34**∗	**41.07**	**27.32**∗∗	26.40∗∗	**25.71**∗∗	**26.48**
Rep src	baseline	35.25	29.69	24.64	29.86	**21.65**	17.40	14.45	17.83
	+TCWR	35.73	**30.48**	25.65	**30.62**	21.57	**17.71**	**14.95**	**18.08**
	+RTT	35.63	30.17	**25.86**	30.55	21.06	17.01	14.36	17.48
	+DRTT(ours)	**35.81**	30.18	25.70	30.56	21.51∗	17.22	14.33	17.69
Rep both	baseline	22.33	18.77	15.98	19.03	25.52	22.68	20.07	22.76
	+TCWR	22.98	19.69	17.14	19.94	25.44	22.64	20.43	22.84
	+RTT	22.92	19.56	16.76	19.75	25.30	22.76	20.66	22.91
	+DRTT(ours)	**23.37**∗∗	**20.23**∗∗	**17.37**∗∗	**20.32**	**26.19**∗	**23.31**∗∗	**20.98**	**23.49**

Table 1: The BLEU scores (%) for forward-translation on noise test sets with noise ratio 0.1, 0.2 and 0.3, and 'AVG' denotes the average BLEU (%) on all noise ratios. We re-implement all baselines to eliminate the discrepancy caused by MLMs and the auxiliary backward model. '∗/∗∗': significantly (Koehn, 2004) better than the RTT with $p < 0.05$ and $p < 0.01$, respectively.

use the baseline model to perform the forward and backward translation in the round-trip translation.

TCWR: Liu et al. (2021) propose the approach of translation-counterfactual word replacement which creates augmented parallel translation corpora by random sampling new source and target phrases from the masked language models.

RTT: Zhang et al. (2021) propose to generate adversarial examples with the single round-trip translation. However, they do not provide any approach for generating the bilingual adversarial pairs. To make a fair comparison, we generate the bilingual adversarial pairs from their adversarial examples in the same way as ours.

5 Results and Analysis

5.1 Main Results

Artificial Noise. To test robustness on noisy inputs, we follow Cheng et al. (2018) to construct five types of synthetic perturbations with different noise ratios on the standard test set[7]: 1) *Deletion:* some words in the source sentence are randomly deleted; 2) *Swap:* some words in the source sentence are randomly swapped with their right neighbors; 3) *Insertion*: some words in the source sentence are randomly repeated; 4) *Rep src:* short for 'replacement on src'. Some words in the source sentence are randomly replaced with their relevant word according to the similarity of word embeddings[8]; 5) *Rep both:* short for 'replacement on both'. Some words in the source sentence and their aligned target words are randomly replaced by masked language models [9].

Table 1 shows the BLEU scores of forward translation results on Zh→En and En→De noise test sets. For Zh→En, our approach achieves the best performance on 4 out of 5 types of noise test sets. Compared to RTT, DRTT achieves the improvement up to 1.1 BLEU points averagely on *deletion*. For En→De, DRTT also performs best results on all types of noise except *Rep src*. We suppose the reason is *Rep src* sometimes reverses the semantics of the original sentence as we claimed above.

Since the perturbations we introduced above may change the semantics of the source sentence, it may be problematic for us to calculate the BLEU score against the original reference sentence in Table 1. Therefore, following Zhang et al. (2021), we also report the BLEU score between the source sentence and its reconstructed version through the source-target-source RTT, which is named as RTT BLEU.

[7]For each test set, we report three results with noise ratio as 0.1, 0.2 and 0.3, respectively. Noise ratio 0.1 means 10 percent of the words in the source sentence are perturbed.

[8]https://github.com/Embedding/Chinese-Word-Vectors
https://nlp.stanford.edu/projects/glove/

[9]Each sentence has four references on NIST test sets, we only choose sb0 for replacement.

Noise	Model	Zh→En				En→De			
		0.1	0.2	0.3	**AVG**	0.1	0.2	0.3	**AVG**
Deletion	baseline	35.31	31.53	28.22	31.69	21.42	19.90	17.42	19.58
	+TCWR	35.02	31.74	28.45	31.74	22.45	20.48	18.66	20.53
	+RTT	35.23	32.12	28.03	31.79	23.34	22.30	20.36	22.00
	+DRTT(ours)	**36.63**$*$	**32.96**$*$	**29.94**$**$	**33.18**	**24.06**$**$	**23.02**$**$	**21.18**$**$	**22.75**
Swap	baseline	28.63	22.82	18.21	23.22	19.01	15.92	14.25	16.39
	+TCWR	31.01	26.03	22.25	26.43	19.56	16.65	14.95	17.05
	+RTT	31.07	26.06	22.08	26.40	20.51	17.63	16.17	18.10
	+DRTT(ours)	**32.03**$*$	**26.95**$**$	**23.71**$**$	**27.56**	**21.40**$**$	**18.68**$**$	**17.53**$**$	**19.20**
Insertion	baseline	30.13	23.57	17.95	23.88	19.57	16.24	13.12	16.31
	+TCWR	30.12	23.76	18.02	23.97	20.73	17.27	**14.12**	17.37
	+RTT	29.72	22.75	17.87	23.45	20.79	16.81	13.80	17.13
	+DRTT(ours)	**31.84**$**$	**24.42**$**$	**19.43**$**$	**25.23**	**21.24**$**$	**17.53**$**$	**14.12**$*$	**17.63**
Rep src	baseline	33.02	28.15	23.26	28.14	20.56	18.40	16.53	18.50
	+TCWR	32.83	28.11	23.38	28.11	21.43	19.22	17.10	19.25
	+RTT	32.65	27.23	23.05	27.65	22.25	20.14	18.45	20.28
	+DRTT(ours)	**34.76**$**$	**29.04**$**$	**25.06**$**$	**29.62**	**22.74**$*$	**20.59**$*$	**18.87**$*$	**20.73**
Rep both	baseline	38.25	36.17	35.48	36.63	23.62	23.23	22.13	22.99
	+TCWR	38.38	36.92	35.44	36.91	24.84	24.77	23.34	24.32
	+RTT	39.13	36.92	35.23	37.09	25.51	24.77	24.12	24.80
	+DRTT(ours)	**40.07**$*$	**38.34**$**$	**37.22**$**$	**38.54**	**26.28**$**$	**25.26**$*$	**24.87**$**$	**25.47**

Table 2: The RTT BLEU scores (%) for round-trip translation on noise test sets. '$*/**$': significantly better than RTT with $p < 0.05$ and $p < 0.01$, respectively.

The intuition behind it is that: a robust NMT model translates noisy inputs well and thus has minor shifting on the round-trip translation, resulting in a high BLEU between inputs and their round-trip translation results. Following Zhang et al. (2021), we fine-tune the backward model with its test set to minimize the impact of the T2S process. As shown in Table 2, DRTT outperforms the other methods on all types of noise on Zh→En and En→De tasks. Considering the results of Table 1 and Table 2 together, DRTT significantly improves the robustness of NMT models under various artificial noises.

Natural Noise. In addition to the artificial noise, we also test the performance of our model on WMT'19 En→Fr robustness test set which contains various noise in real-world text, e.g., exhibits typos, grammar errors, code-switching, etc. As shown in Table 3, DRTT yields improvements of 1.25 BLEU compared to the baseline, it proves that our approach also performs well in real noise scenario. Besides, DRTT achieves 0.63 BLEU improvement over RTT by filtering out 10% of fake adversarial examples (according to Table 6 in Appendix B), which demonstrates that filtering out fake adversarial examples further improves the robustness of the model.

5.2 Effectiveness of Adversarial Examples

In this sub-section, we evaluate the effectiveness of the generated adversarial examples on attacking the victim NMT model (i.e., the target NMT model

Method	**En→Fr**	**BLEUΔ**
baseline	35.11	
+TCWR	35.64	+0.53
+RTT	35.73	+0.62
+DRTT(ours)	**36.36**$*$	**+1.25**

Table 3: The BLEU scores (%) on the WMT'19 En→Fr robustness task. 'BLEUΔ' denotes the gain of BLEU compared to baseline. '$*/**$': significantly better than RTT with $p < 0.05$ and $p < 0.01$, respectively.

without being trained on the generated adversarial pairs). In our approach, γ in Eq.(4) is a hyper-parameter to control the strictness of our definition on generating adversarial examples. Thus, we evaluate the effectiveness of adversarial examples by studying the translation performance of the victim NMT model on the set of adversarial pairs generated with different γ. That is to say, if a sample is an adversary, it should destroy the translation performance drastically, resulting in a low BLEU score between the translation result and its paired target sentence. The average BLEU scores of the victim model on the different adversarial pair sets (generated with γ from -10 to 1 on NIST 06) are shown in Figure 3. Specifically, the average BLEU on the adversarial sets generated with $\gamma = -10$ is 8.0. When we remove the restriction of γ, i.e., the DRTT is degenerated into RTT, the average BLEU for the constructed adversarial examples reaches up to 11.2. This shows that the adversarial examples generated with lower γ (more strict restriction)

Model	Zh→En							En→De	
	MT06	MT02	MT03	MT04	MT05	MT08	**AVG**	newstest16	newstest17
baseline	44.59	44.38	43.65	45.37	44.42	35.80	42.72	29.11	27.94
+TCWR	44.55	**45.99**	44.68	45.77	44.16	34.98	43.12	29.13	27.98
+RTT	44.62	45.13	44.01	46.00	**44.96**	35.18	43.06	29.06	27.42
+DRTT(ours)	**44.76**	45.01	**45.16**$**$	**46.63**$**$	44.78	**35.82**$*$	**43.48**	**29.30**	**28.37**$**$

Table 4: The BLEU scores (%) on NIST Zh→En and WMT17 En→De. '$*$/ $**$': significantly better than RTT with $p < 0.05$ and $p < 0.01$, respectively.

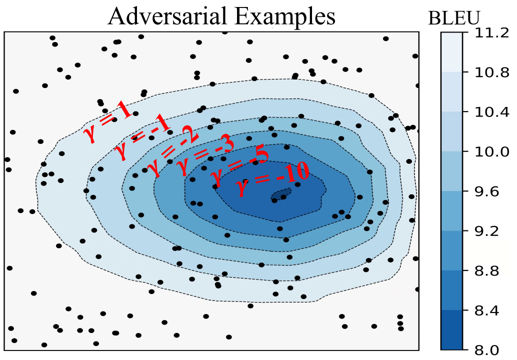

Figure 3: Black spots represent the distribution of adversarial samples. The darker color indicates more effective adversarial examples generated with lower γ.

x : 我们坚决拥护政府处理这一事件所采取的措施。 y : we resolutely support measures taken by our government in handling this incident.
x_δ : 我们坚决反对政府处理这一案件所采取的举措。 y_δ : we resolutely oppose measures taken by our government in handling this case.
x : 中美双方认为, 当前世界经济形势是良好的。通货膨胀继续保持低水平, 大多数新兴市场经济体的经济增长强劲。 y : china and the united states agreed that the present economic situation in the world is satisfactory, with inflation kept at a low level and most of the new market economies growing strong.
x_δ : 俄美双方认为, 当前世界贸易势头是不错的。通货膨胀继续保持低速度, 大多数新兴市场经济体的经济发展强劲。 y_δ : russia and the united states agreed that the present trade trend in the world is satisfactory, with inflation kept at a low rate and most of the new market economies developing strong.

Table 5: Case study for the proposed approach. The words in red and blue color represents the augmented words on the source and target side, respectively.

attack the model more successfully. Therefore, We can select more effective adversarial examples compared to Zhang et al. (2021) by lowering the threshold γ to create a more strict definition.

5.3 Clean Test set

Adding a large amount of noisy parallel data to clean training data may harm the NMT model performance on the clean test sets seriously (Khayrallah and Koehn, 2018). In this sub-section, we test the performance of the proposed model on the clean test sets and the results are presented in Table 4. DRTT achieves the best translation performance on Zh→En and En→De clean test sets. It demonstrates that our approach not only improves the robustness of the NMT model, but also maintains its good performance on clean test sets.

5.4 Case Study

In Table 5, we present some cases from Zh-En adversarial pairs generated by our approach. From the case 1, we can see "拥护" in the source sentence is replaced by its antonym "反对", which reverse the meaning of the original sentence, and DRTT makes a corresponding change in the target sentence by replacing "support" with "oppose". In the other case, DRTT replaces "良好" by its synonym "不错", thus, "satisfactory" in the target sentence remains unchanged. From these cases, we find that DRTT can reasonably substitute words in source sequences based on the contexts and correctly modify the corresponding target words synchronously.

6 Conclusion and Future Work

We propose a new definition for NMT adversarial examples based on Doubly Round-Trip Translation, which can ensure the examples that meet our definition are the authentic adversarial examples for the target NMT model. Additionally, based on this definition, we introduce the masked language models to generate bilingual adversarial pairs, which can be used to improve the robustness of the NMT model substantially. Extensive experiments on both the clean and noise test sets (including artificial and nature noise) show that our approach not only improves the robustness of the NMT model but also performs well on the clean test sets. We will explore improving the robustness of target and backward models simultaneously in the future.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Guandan Chen, Kai Fan, Kaibo Zhang, Boxing Chen, and Zhongqiang Huang. 2021. Manifold adversarial augmentation for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3184–3189.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333.

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

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. *arXiv preprint arXiv:1805.06130*.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation 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 187–197.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2021. Attention calibration for transformer in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1288–1298, Online. Association for Computational Linguistics.

Paul Michel, Xian Li, Graham Neubig, and Juan Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3103–3114.

Paul Michel and Graham Neubig. 2018. Mtnt: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553.

Xing Niu, Prashant Mathur, Georgiana f Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. 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. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. 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.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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.

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.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. Csp: Code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977, Online. Association for Computational Linguistics.

A Bilingual Adversarial Pair Generation

Algorithm 1: Bilingual Adversarial Pair Generation

Input: A sequence pair (x, y), a sampling probability c, an alignment mapping p, candidate words k, masked language models M-MLM and T-MLM, thresholds β and γ.

Output: A bilingual adversarial pair (x_δ, y_δ)

Function `BilAdvGen` (x, y)**:**
  while $i \leq len(x) * c$ **do**
    $r_i^j \leftarrow$ M-MLM $(x^{\backslash i})$;
    $x^{\backslash i:j} \leftarrow$ Replace(x, r_i^j)
    $r_i^* \leftarrow \arg\max \mathrm{d_{src}}(x, x^{\backslash i:j})$ (2);
    $x^{\backslash i:*} \leftarrow$ Replace(x, r_i^*)
    Get aligned index $p(i)$;
    $w_{p(i)} \leftarrow$ T-MLM $(x^{\backslash i:*}, y^{\backslash p(i)})$;
    $y_\delta \leftarrow$ Replace$(y, w_{p(i)})$
    $i \leftarrow i + 1$;
  end
if $\mathrm{d_{src}}(x, x_\delta) > \beta$ and $\mathrm{d_{tgt}}(y, y_\delta) < \gamma$ **then**
  return x_δ, y_δ
end

B Implementation Details

As for Zh→En, we apply the separate byte-pair encoding (BPE) (Sennrich et al., 2016) encoding with 30K merge operations for Zh and En, respectively, the peak learning rate of 5e-4, and the training step is 100K. For En→De and En→Fr, we apply the joint BPE with 32K merge operations, the learning rate of 7e-4 and the training step is 200K. The dropout ratio is 0.1. We use Adam optimizer (Kingma and Ba, 2014) with 4k warm-up steps.

Method	**Zh→En**	**En→De**	**En→Fr**
original	1252977	5859951	2037962
-RTT	1236485	2670044	1639661
-DRTT(ours)	956308	2336285	1466756

Table 6: The statistics of the number of adversarial examples generated by different method. RTT denotes generated with Eq.(2). DRTT denotes generated with Eq.(4).