APPENDIX

In this appendix, we provide necessary background of Neural Machine Translation (NMT), pre-trained language models, and the back-translation technique used in related works. Besides, screenshots of Table 8 are also provided.

Neural Machine Translation. Typical NMT models follow an encoder-decoder architecture with attention mechanisms (Zhang et al., 2019). The encoder encodes the source language to a latent representation space, and the decoder is a neural language model that decodes representations in the latent space to another language domain. Either the encoder or the decoder can be built on recurrent neural networks (Bahdanau et al., 2015), convolutional neural networks (Costa-jussà and Fonollosa, 2016), or Transformer networks (Vaswani et al., 2017). In this work, we applied two versions of neural network architecture for the encoder/decoder models: RNN and Transformer.

Pre-trained Language Model. Recently, pre-trained language models, such as mask language models (MLM) (Devlin et al., 2019), have achieved a powerful initialization for the NMT encoder models. MLM pre-trains the encoder for a better language understanding on the encoded language by randomly masking some tokens in continuous monolingual text streams and predicting these tokens. To predict the masked tokens, the language model pays attention to the relative language parts, which encourages the model to have a better understanding on the language. Inspired by the powerful language understanding ability of the pre-trained language models, and following the black-box setting, we use the pre-trained MLM to estimate the word saliency and build the word embedding space for adversarial attacks.

Back-Translation. There are a lot of works for improving the NMT performance by leveraging the back translation, which uses not only parallel corpus but also monolingual corpus for training the NMT models (He et al., 2016; Lample and Conneau, 2019). Previous works on back-translation demonstrate the ability of the dual NMT models to reconstruct the language. In this work, we observe that the back-translation technique makes it possible to evaluate NMT adversarial attacks without ground-truth references for the perturbed sentences, and we propose to evaluate the proposed NMT attack method basing on the reconstruction results of the original inputs and the perturbed examples.

Figure 3: An example of attacking the baidu translator, in which the adversarial example is generated on the *Rnns.* model using WSLS.

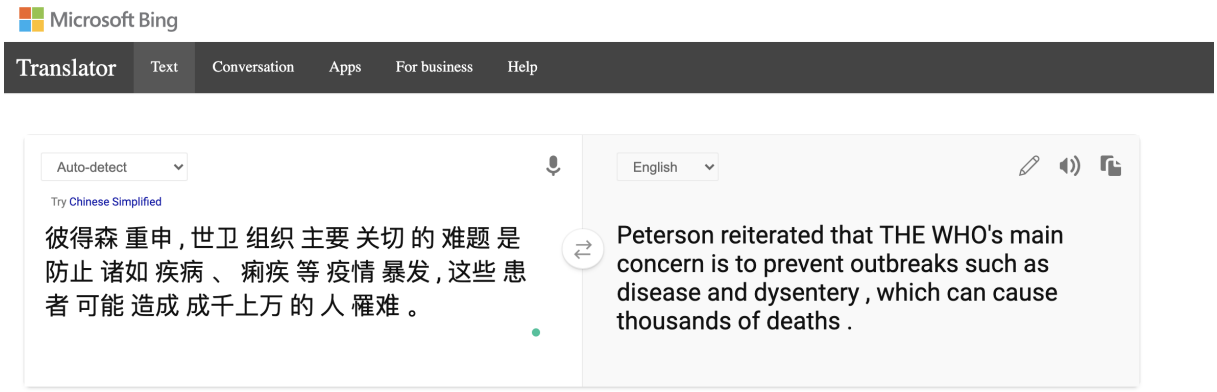

Figure 4: An example of attacking bing translator, in which the adversarial example is generated on the *Rnns.* model using WSLS.