LOST IN TRANSLATION: GENERATING ADVERSARIAL EXAMPLES ROBUST TO ROUND-TRIP TRANSLATION

Neel Bhandari
RV College of Engineering, India
neelbhandari.cs18@rvce.edu.in

Pin-Yu Chen
IBM Research
pin-yu.chen@ibm.com

ABSTRACT

Language Models today provide a high accuracy across a large number of downstream tasks. However, they remain susceptible to adversarial attacks, particularly against those where the adversarial examples maintain considerable similarity to the original text. Given the multilingual nature of text, the effectiveness of adversarial examples across translations and how machine translations can improve the robustness of adversarial examples remain largely unexplored. In this paper, we present a comprehensive study on the robustness of current text adversarial attacks to round-trip translation. We demonstrate that 6 state-of-the-art text-based adversarial attacks do not maintain their efficacy after round-trip translation. Furthermore, we introduce an intervention-based solution to this problem, by integrating Machine Translation into the process of adversarial example generation and demonstrating an increased robustness to round-trip translation. Our results indicate that finding adversarial examples robust to round-trip translation can help identify insufficiency of language models that is common across languages, and motivate further research into multilingual adversarial attacks.

1 INTRODUCTION

Language models, despite their remarkable success across tasks, have shown to be vulnerable to adversarial examples, which are inputs designed to be similar to the model's native data inputs, but crafted with small modifications to fool the model during inference. These examples can be classified correctly by a human observer, but often mislead a target model, providing an insight into their robustness to adversarial inputs.They are essential in understanding key vulnerabilities in models across a variety of applications. ML models are being increasingly deployed commercially for translation. A special form of translation is round trip translation, which focuses on translating a given text from one language to the second and back to the first. Round trip translation has been increasingly used in several research areas, including correcting grammatical errors Lichtarge et al. (2019); Madnani et al. (2012), evaluating machine translation models Crone et al. (2021); Cao et al. (2020); Moon et al. (2020); Shigenobu (2007), paraphrasing Guo et al. (2021) and rewriting questions Chu et al. (2019). It is also used extensively as part of the quality assurance process in critical domains such as medical, legal and market search domains, where they use back-translations. The use of ML models in these critical domains means that they have to be tested by robust adversarial attacks to make for safe commercial deployment. Given the importance of round trip translation, we are motivated to study its effects on current adversarial attacks.

Contributions. (i) We demonstrate that round trip translation can be used as a cheap and effective defense against *current* textual adversarial attacks. We show that 6 state-of-the-art adversarial text attacks suffer an average performance loss of 66%, rendering most examples generated non-adversarial. (ii) However, we find that round-trip translation defensive capabilities can be bypassed by our proposed novel *attack-agnostic* algorithm that provides machine translation intervention to increase robustness against round-trip translation. We also find that there is minimal difference in quantification metrics to the original, which shows our method finds a new set of robust and high-quality text adversarial examples against NMT.

Related works. Papernot et al. (2017) proposed a white box adversarial attack that repeatedly modified the input text till the generated text fooled the classifier. This method, although effective in

Algorithm 1: NMT-Text-Attack

Input : Sentence $S = [w_1, w_2, .., w_n]$, Ground truth label Y, Victim Model V, Machine Translation model M, User-Specific Constraints C, Attack A

Output: Adversarial Example X_{adv}

Phase I - Word Importance Ranking
Call attack A & Initialize edge weights
for *each word* w_i *in* S **do**
    Compute Importance score I_i from A
Sort words in descending order into list W
Phase II - Word Replacement
# Word Replacement Strategy
for *each word* w_i *in* W **do**
    Predict Top-K replacements for w_i using A and store in $R = [r_1, r_2, .., r_k]$
    for *each word* w_i *in* W **do**
        Replace w_i with r_j in S to make X_{adv}
        Round-Trip-Translate X_{adv} with P language(s) using M to make $T = [t_1, t_2, .., t_p]$ where t_i is X_{adv} translated through language i
        Evaluate classification scores for $T = [t_1, t_2, .., t_p]$ using V, removing examples that do not maintain adversarial sentiment
        for *each* $c_i \in C$ **do**
            Apply constraint c_i to each $t_i \in T$
        Select best $t_i \in T$ w.r.t constraints C and store as X_{adv}
return X_{adv}

principle, did not maintain semantic meaning of the sentence. Ebrahimi et al. (2018) as well as Samanta & Mehta (2017) gradient-based solutions involving token based changes and searching for important words. These methods, however, did not prove to be scalable and lacked robust performance. It was followed by methods such as character replacement Ribeiro et al. (2018), phrase replacement and word scrambling. These techniques, however, fail to maintain semantic consistency with the original input. Jia et al. (2019) introduced adding distracting sentences to the reading comprehension task.Jin et al. (2020) propose TextFooler which generates adversaries using token level similarity and bound by axiomatic constraints. Garg & Ramakrishnan (2020) introduce BAE, which uses masked-language modelling to generate natural adversarial examples for the text. Recent works in adversarial attacks on NMT include Cheng et al. (2019) using gradient based adversarial inputs to improve robustness of NMT models, and Zhang et al. (2021) proposed a novel black-box attack algorithm for NMT systems. However, none of these works target round-trip translation, and do not demonstrate attack agnostic capabilities.

2 NMT-TEXT-ATTACK: PROPOSED METHOD

In order to generate adversarial examples robust to round-trip translation, we propose an intervention-based, attack-agnostic method that only requires access to a neural machine translation model. We employ a generic template used by standard state-of-the-art adversarial attack examples in order to showcase the attack-agnostic capabilities. From Li et al. (2019); Jin et al. (2020); Ren et al. (2019); Garg & Ramakrishnan (2020); Gao et al. (2018) it can be seen that the attacks follow a two section split. The first section is word importance ranking, where importance scores are computed and sorted, and the second section deals with word replacement, where attack specific constraints are attached. The second section is where we introduce our NMT model as an added constraint for round trip translation and introduce NMT-Text-Attack.

I. Word Importance Selection. This section initially involves pre-processing the input sentence with techniques such as removing stop words, followed by analysing the most important keywords in the target sentence using several techniques, ranging from the input deletion method, to probability weighted word saliency. These methods are specific to the adversarial attack chosen to be integrated with NMT-Text-Attack. For example, TextFooler uses the input deletion method.. Once the most

important words are learnt, attack algorithms look for replacements through synonym search or by replacing individual characters of the original input word to make an adversarial candidate.

II. Constraint Evaluation. This section involves introducing constraints to maximise the desired performance of the algorithm. Examples of such constraints are semantic similarity to original input on replacement, POS tag preservation etc. Again, these are specific to the algorithm being implemented. We introduce the machine translation task in this section. We collect the candidate sentences from the word importance selection and run them through the neural machine translation model. We implement round-trip translation on these sentences for k languages, where k is specified by the user . We then evaluate the translated sentences on the task, and only select those translated sentences which provide confident results in the adversarial attack. The final adversarial example is then selected by analysing which of the final candidate examples had the highest similarity score with respect to the original sentence.

3 Performance Evaluation

For evaluation, we use a range of algorithms from the TextAttack library Morris et al. (2020). We have listed all of the algorithms used in the Appendix, along with the experimental details.

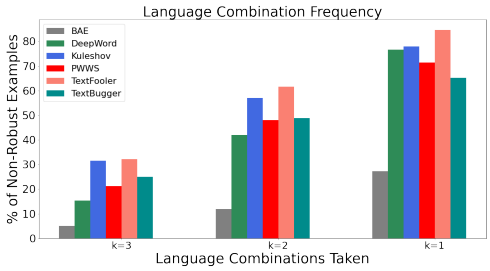
Dataset and Victim Model We use the Rotten Tomatoes Movie Reviews Pang & Lee (2005) and Yelp Polarity Zhang et al. (2015) datasets to perform sentiment analysis. The datasets consist of an equal split in sentiments. We sample 1000 random examples from the test set of each of these mentioned datasets and run our experiments on them.
For our Victim Model, we use the BERT model Devlin et al. (2019), pre-trained on the specific datasets mentioned from the HuggingFace Library Wolf et al. (2020).

Current Attacks are not Robust to Round Trip Translation. We run 6 adversarial attacks on the Movie Reviews Dataset and analyse their robustness to round-trip translation. We analyse them against 3 languages- Spanish, German and French through the EasyNMT library Tang et al. (2020) (see Appendix for more details). On round-trip translating the adversarial examples, we test the resultant examples against the classification model. On the y-axis, we provide the percentage of non-robust examples to at least k out of $m = 3$ languages. Formally, if k is the number of languages used in tandem, N is the number of examples in total, y_a is the original prediction before round trip translation and $\hat{y_a}$ is the prediction after round-trip translation by translation model M and victim model V, then the y-axis is defined as $Y = \frac{1}{N}\sum_{a=1}^{N} \mathbb{1}\{\text{at least k languages have } y_a \neq \hat{y_a}\}$, where $\mathbb{1}\{E\}$ is an indicator function such that it is one when the event E is true and zero otherwise.
We see that on average, over 66% of the examples generated originally by the attack are rendered non-adversarial on round-trip translation with at least one language ($k = 1$). BAE remains the most robust to translations, while TextFooler remains the least robust. On increasing the number of language combinations taken ($k > 1$), we see that there is a decrease in effectiveness of round trip translation as a defense against the adversarial examples, however there is still significant loss in attack success rate. This is because when you add more languages as a constraint, there is an increased chance that at least one of the constrained languages is robust to round-trip translation for any example. This provides considerable evidence that round trip translation can be used as a cheap and effective defense, and motivates the question of whether there exists text adversarial examples robust to round-trip translation. In the following sections, we detail a simple, attack-agnostic algorithm that shows that such examples exist and can be used readily by current attacks.

Table 1: Success Rate (%) of NMT-Text-Attack Relative to when Original Attack Success Rate is 100%

Dataset	TextFooler	TextBugger	PWWS
MR	70.7	74.7	69.4
Yelp	60.0	71.4	68.8

Figure 1: Percentage of non-robust examples flagged by at least k language combination



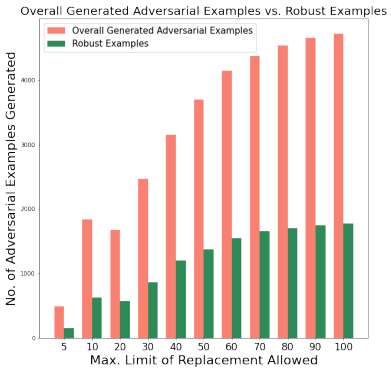
NMT-Text-Attack Results. We analyse the results of incorporating NMT-Text-Attack into existing attacks across the mentioned datasets. We evaluate the attack on its success rate with respect to the attacks' native success rate without NMT-Text-Attack. Note that, through our novel intervention-based algorithm, we are able to guarantee 100% robustness to back-translation on the user's selected language(s). This is because our algorithm (line 14) introduces a strict constraint to only allow examples that are robust to back-translation to be selected as candidates for the attack, which leads to significant increase over the original algorithm's robustness to round-trip translation. This guarantee is important as it helps achieve high-quality robustness in multilingual settings, which no existing adversarial attack can provide. Table 1 shows that to meet this criteria, NMT-Text-Attack is successful on average 30% less examples than its original counterpart.

While this loss may seem significant, we believe this is justified for two reasons. First, this loss comes with a 100% success in robustness to round-trip translation coupled with attack success. This is critical in commercial settings where deployed models need to have confident outputs in the face of several language translations. Secondly, in Figure 2, we see that there is considerable scope to increase the number of robust examples available simply by increasing the replacement limit. We set our replacement limit as mentioned in the Appendix for our experiments, and Figure 2 demonstrates that scaling the number of replacements significantly increases number of available robust examples.

Table 2: Performance of NMT-Text-Attack on Yelp and Movie Reviews (MR) Datasets

Dataset	Attack	USE	Jaccard	BERT
Yelp	TextBugger + NMT	0.94	0.848	0.9715
	TextFooler + NMT	0.82	0.724	0.956
	PWWS + NMT	0.83	0.645	0.9265
	TextBugger	0.93	0.79	0.95
	TextFooler	0.93	0.81	0.97
	PWWS	0.93	0.85	0.97
MR	TextBugger + NMT	0.91	0.68	0.92
	TextFooler + NMT	0.82	0.724	0.956
	PWWS + NMT	0.83	0.645	0.9256
	TextBugger	0.93	0.79	0.95
	TextFooler	0.813	0.715	0.953
	PWWS	0.85	0.77	0.96

Figure 2: Replacement vs. Robust Examples



We also provide a quantitative analysis of our model by analysing the adversarial examples generated against the original attack in Table 2. Universal Sentence Encoder with cosine similarity, along with Jaccard Similarity are used as similarity metrics, while BERT Score is used to analyze meaning preservation. We notice that there is little variation in the effectiveness of the algorithms when it comes to meaning preservation and similarity, which shows that our proposed intervention, while increasing robustness significantly, maintains the quality of the original attack. Examples of adversarial examples on sentences have been mentioned in the Appendix.

In addition to these results, we have provided an ablation study that tests our algorithm on unseen languages. Compared to the original attacks, we see a 20% increase in robustness to unseen languages. The ablation study can be found in Appendix (Section 6.4).

4 Conclusion

In this paper, we demonstrate the ineffectiveness of current text adversarial attack algorithms to round-trip translation, and provide an intervention-based method to improve robustness to round-trip translation in these algorithms. We show that this intervention (NMT-Text-Attack) has minimal effect on the actual semantic metrics but can significantly improve the attack success rate against back-translation, suggesting that there exist a new set of robust text adversarial examples.The attack-agnostic nature of the algorithm along with its high-quality performance makes it an effective error diagnosing tool with any existing text attack for inspecting model robustness.

REFERENCES

Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ, 1972.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples, 2018.

American Psychological Association. *Publications Manual*. American Psychological Association, Washington, DC, 1983.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, December 2005. ISSN 1532-4435.

Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 33–40, 2007.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015.

Jialun Cao, Meiziniu Li, Yeting Li, Ming Wen, and S. C. Cheung. Semmt: A semantic-based testing approach for machine translation systems. *ArXiv*, abs/2012.01815, 2020.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

Sin-wai Chan. *A Dictionary of Translation Technology*. The Chinese University of Hong Kong Press, 2006.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133, 1981. doi: 10.1145/322234.322243.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples, 2020.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. Robust neural machine translation with doubly adversarial inputs, 2019.

Zewei Chu, Mingda Chen, Jing Chen, Miaosen Wang, Kevin Gimpel, Manaal Faruqui, and Xiance Si. How to ask better questions? a large-scale multi-domain dataset for rewriting ill-formed questions, 2019.

Nathan E. Crone, A. J. Power, and John Weldon. Quality estimation using round-trip translation with sentence embeddings. *ArXiv*, abs/2111.00554, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 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)*, pp. 31–36, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2006. URL https://aclanthology.org/P18-2006.

Shi Feng, Eric Wallace, Alvin Grissom II au2, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult, 2018.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers, 2018.

Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

Zilu Guo, Zhongqiang Huang, Kenny Q. Zhu, Guandan Chen, Kaibo Zhang, Boxing Chen, and Fei Huang. Automatically paraphrasing via sentence reconstruction and round-trip translation. In *IJCAI*, 2021.

Dan Gusfield. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK, 1997.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions, 2019.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment, 2020.

Volodymyr Kuleshov, Shantanu Thakoor, Tingfung Lau, and Stefano Ermon. Adversarial examples for natural language classification problems. 2018.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized perturbation for textual adversarial attack, 2021.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *ArXiv*, abs/1812.05271, 2019.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure, 2017.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert, 2020.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Jul 2018. doi: 10.24963/ijcai.2018/585. URL http://dx.doi.org/10.24963/ijcai.2018/585.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. Corpora generation for grammatical error correction. 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)*, pp. 3291–3301, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1333. URL https://aclanthology.org/N19-1333.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. Exploring grammatical error correction with not-so-crummy machine translation. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 44–53, Montréal, Canada, June 2012. Association for Computational Linguistics. URL https://aclanthology.org/W12-2005.

George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. Revisiting round-trip translation for quality estimation, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations, 2017.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, 2020.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of HLT-NAACL*, 2016.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints, 2016a.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints, 2016b.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning, 2017.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Mohammad Sadegh Rasooli and Joel R. Tetreault. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733, 2015. URL http://arxiv.org/abs/1503.06733. version 2.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1085–1097, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1103. URL https://aclanthology.org/P19-1103.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1079. URL https://aclanthology.org/P18-1079.

Suranjana Samanta and Sameep Mehta. Towards crafting text adversarial samples, 2017.

Tomohiro Shigenobu. Evaluation and usability of back translation for intercultural communication. In Nuray Aykin (ed.), *Usability and Internationalization. Global and Local User Interfaces*, pp. 259–265, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-73289-1.

Harold Somers. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pp. 127–133, Sydney, Australia, December 2005. URL https://aclanthology.org/U05-1019.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning, 2020.

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. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. Word-level textual adversarial attacking as combinatorial optimization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.540. URL http://dx.doi.org/10.18653/v1/2020.acl-main.540.

Zhiyuan Zeng and Deyi Xiong. An empirical study on adversarial attack on NMT: Languages and positions matter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 454–460, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.58. URL https://aclanthology.org/2021.acl-short.58.

Zhiyuan Zeng and Deyi Xiong. An empirical study on adversarial attack on nmt: Languages and positions matter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 454–460, 2021b.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification . *arXiv:1509.01626 [cs]*, September 2015.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 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)*, pp. 1967–1977, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.153. URL https://aclanthology.org/2021.acl-long.153.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples, 2018.

A Appendix

Ethical Concerns

Our paper discusses the potential weakness of NLP models to round-trip translation, and describes an algorithm that can make the weakness more robust. However, we believe that we give new insights in studying text adversarial examples and will spur more robust machine learning models in the future. We are also the first individuals to introduce the vulnerability to round-trip translation, which provides opportunity to develop robust models in a novel setting.

A.1 Computational Resources

For the implementation of our algorithm and experiments, we use Google Colab as our base GPU provider. The GPU typically provided is is Tesla - P100. We use 190 GPU hours to run all our experiments. We use a pre-trained BERT model with 12-head attention and 110 million parameters, which is typical of BERT models.

A.2 Machine Translation Setup

We use the Opus-MT set of models through the EasyNMT library Tang et al. (2020). Opus-MT consists of 1200 models trained on several languages for open translation. The architecture for the Opus-MT models is based on a standard transformer setup with 6 self-attentive layers in both, the encoder and decoder network with 8 attention heads in each layer. This architecture is used to back-translate the target reviews from English to French, German and Spanish, and back to English.

A.3 ADVERSARIAL ATTACK SETTINGS

Algorithm 1 details a general template of several state of the art adversarial attacks we have used in the paper. In this section we detail the exact settings used for each adversarial attack when integrated with NMT-Text-Attack. These are standard approaches used directly from the TextAttack Library with no changes in standard settings.

A.3.1 TEXTFOOLER

- Word Importance Selection
  - Max allowable replacement candidate generation for synonyms: 40.
  - Transformation Embedding Mechanism: Counterfitted Glove Embeddings Mrkšić et al. (2016a)
- Word Replacement:
  - Pre-transformation constraints:
    - RepeatModification: A constraint disallowing the modification of words which have already been modified
    - StopwordModification: A constraint disallowing the modification of stopwords
  - Constraints:
    - Minimum cosine distance between word embeddings = 0.5
    - Part of Speech : Only replace words with the same part of speech (or nouns with verbs)
    - Universal Sentence Encoder with a minimum angular similarity of = 0.5.
    - Word Swapping Technique: Greedy Word Swap with Word Importance Ranking with word importance ranking conducted using input deletion method.

A.3.2 TEXTBUGGER

- Word Importance Selection
  - Max allowable replacement candidate generation for synonyms: 40.
  - Transformation Embedding Mechanism: Counterfitted Glove Embeddings Mrkšić et al. (2016a)
  - Allowable Swap Mechanisms: Character Insertion, Character Deletion, Adjacent Character Swap, Homoglyph Swap.
- Word Replacement:
  - Pre-transformation constraints:
    - RepeatModification: A constraint disallowing the modification of words which have already been modified
    - StopwordModification: A constraint disallowing the modification of stopwords
  - Constraints:
    - Universal Sentence Encoder with a minimum angular similarity of = 0.84
    - Word Swapping Technique: Greedy Word Swap with Word Importance Ranking with word importance ranking conducted using input deletion method.

A.3.3 PWWS

- Word Importance Selection
  - Max allowable replacement candidate generation for synonyms: 40.
  - Transformation Embedding Mechanism: Word Swap by swapping synonyms in WordNet Miller (1998)
  - Allowable Swap Mechanisms: Character Insertion, Character Deletion, Adjacent Character Swap, Homoglyph Swap.
- Word Replacement:
  - Pre-transformation constraints:

        * RepeatModification: A constraint disallowing the modification of words which have already been modified
        * StopwordModification: A constraint disallowing the modification of stopwords
    - Constraints:
        * Word Swapping Technique: Greedy Word Swap with Word Importance Ranking with word importance ranking conducted using weighted saliency method.

A.3.4 KULESHOV

- Word Importance Selection
    - Max allowable replacement candidate generation for synonyms: 15.
    - Transformation Embedding Mechanism: Counterfitted Glove Embeddings Mrkšić et al. (2016a)
- Word Replacement:
    - Pre-transformation constraints:
        * RepeatModification: A constraint disallowing the modification of words which have already been modified
        * StopwordModification: A constraint disallowing the modification of stopwords
    - Constraints:
        * Max words perturbed = 50
        * Maximum thought vector Euclidean distance = 0.2
        * Maximum language model log-probability difference = 2
        * Word Swapping Technique: Greedy Word Search.

A.3.5 DEEPWORDBUG

- Word Importance Selection
    - Max allowable replacement candidate generation for synonyms: 40
    - Embedding Transformation Mechanism: Counterfitted Glove Embeddings Mrkšić et al. (2016a)
    - Allowable Swap Mechanisms: Character Insertion, Character Deletion, Adjacent Character Swap, Random Character Substitution.
- Word Replacement:
    - Pre-transformation constraints:
        * RepeatModification: A constraint disallowing the modification of words which have already been modified
        * StopwordModification: A constraint disallowing the modification of stopwords
    - Constraints:
        * Maximum Levenshtien Edit Distance= 30.
        * Word Swapping Technique: Greedy Word Swap with Word Importance Ranking with word importance ranking conducted using input deletion method.

A.3.6 BAE

- Word Importance Selection
    - Max allowable replacement candidate generation for synonyms: 40
    - Transformation Embedding Mechanism: Transformer AutoTokenizer and word replacement using Masked Language Modelling. Mrkšić et al. (2016a)
- Word Replacement:
    - Pre-transformation constraints:
        * RepeatModification: A constraint disallowing the modification of words which have already been modified
        * StopwordModification: A constraint disallowing the modification of stopwords
    - Constraints:

* Part of Speech : Only replace words with the same part of speech (or nouns with verbs)
* Universal Sentence Encoder with a minimum angular similarity = 0.93.
* Word Swapping Technique: Greedy Word Swap with Word Importance Ranking with word importance ranking conducted using input deletion method.

A.4 Ablation Study

In this section, we provide an ablation study to substantiate the performance of our algorithm. In this study, we provide TextFooler with NMT-Text-Attack with 2 'seen' languages and test its performance with an 'unseen' language. A 'seen' language is defined as one which model is provided with as constraints for adversarial examples to satisfy,as shown in Algorithm 1. An 'unseen' language, consequently, is one which the model has not added as a constraint, hence does not guarantee 100% robustness against. The three languages we use are French, German, and Spanish. We alternate between using two of the languages as 'seen', and one as 'unseen'. We compare this with the performance of TextFooler without NMT-Text-Attack on the unseen languages in Table 3. We observe that TextFooler with NMT-Text-Attack outperforms TextFooler without NMT-Text-Attack on average by 20%. This shows that the integration of our attack-agnostic algorithm provides significant performance increase even in situations where the attack is facing unseen languages.

A.5 Examples of NMT-TextAttack

1. **Original** : drawing on an irresistible , languid romanticism , byler reveals the ways in which a sultry evening or a beer-fueled afternoon in the sun can inspire even the most retiring heart to venture forth . **(Sentiment: Positive)**

Adversarial (TextFooler): drawing on an gargantuan , lolling melodrama , byler betrays the ways in which a sultry evening or a beer-fueled afternoon in the sun can inspire even the most retiring heart to venture forth . **(Sentiment: Negative)**

Adversarial (TextFooler+NMT-Text-Attack): drawing on an inexorable, crooning melodrama byler reveals the ways in which a sultry evening or a beer-fueled afternoon in the sun can inspire even the most retiring heart to venture forth. **(Sentiment: Negative)**

Back-Translated (TextFooler): drawing on a giant melodrama, melodrama lolling, Byler betrays the ways in which a sensual afternoon or an afternoon of beer fed in the sun can inspire even the most outgoing heart to venture forward. **(Sentiment: Positive)**

Back-Translated (TextFooler+NMT-Text-Attack): drawing on a melodrama byler inexorable betrays the ways in which a sensual afternoon or an afternoon of beer fed in the sun can inspire even the most outgoing heart to venture forward **(Sentiment: Negative)**

2. **Original** : Exceptionally well acted by Diane Lane and Richard Gere . **(Sentiment: Positive)**

Adversarial (TextFooler): Exceptionally opportune acted by Diane Lane and Richard Gere .**(Sentiment: Negative)**

Adversarial (TextFooler+NMT-Text-Attack): Exceptionally better acted by Diane Lane and Richard Gere **(Sentiment: Negative)**

Back-Translated (TextFooler): exceptionally timely performed by Diane Lane and Richard Gere. **(Sentiment: Positive)**

Table 3: Performance of NMT-Text-Attack on unseen language

Seen Languages	Unseen Language	TextFooler +NMT	TextFooler w/o NMT
French and German	Spanish	72.9%	50.61
French and Spanish	German	74.08%	51.97
German and Spanish	French	67%	50.8

Back-Translated (TextFooler+NMT-Text-Attack): exceptionally better performed by Diane Lane and Richard Gere **(Sentiment: Negative)**

3.**Original** : this kind of hands-on storytelling is ultimately what makes shanghai ghetto move beyond a good , dry , reliable textbook and what allows it to rank with its worthy predecessors . **(Sentiment: Positive)**

Adversarial (PWWS): this tolerant of hands-on storytelling is ultimately what piss shanghai ghetto move beyond a good , dry , reliable textbook and what allows it to gross with its worthy predecessors **(Sentiment: Negative)**

Adversarial (PWWS+NMT-TextAttack):this tolerant of hands-on storytelling is ultimately what makes shanghai ghetto move beyond a good , dry , reliable textbook and what allows it to place with its worthy predecessors . **(Sentiment: Negative)**

Back-Translated (PWWS): This tolerant of practical narration is ultimately what pis shanghai ghetto move beyond a good, dry, reliable textbook and what allows rough with its worthy predecessors. **(Sentiment: Positive)**

Back-Translated (PWWS+NMT-Text-Attack): this tolerant of narration is ultimately what builds the shanghai ghetto to move beyond a good reliable dry text book and what allows it to grossly with its worthy predecessors. **(Sentiment: Negative)**

4.**Original** : I went there today! I have an awful experience. They lady that cut my hair was nice but she wanted to leave early so she made a disaster in my head! **(Sentiment: Positive)**

Adversarial (PWWS): I went there today! I have an awesome experience. They lady that cut my hair was nice but she wanted to leave early so she made a disaster in my head!**(Sentiment: Negative)**

Adversarial (PWWS+NMT-TextAttack):I went there today! I have an direful experience! They lady that cut my hair was nice but she wanted to leave early so she made a disaster in my head **(Sentiment: Negative)**

Back-Translated (PWWS): I went there today. I have a amazing experience. The lady who cut my hair was nice, but she wanted to leave early, so she made a mess of my head. **(Sentiment: Positive)**

Back-Translated (PWWS+NMT-Text-Attack): I went there today. I have a terrible experience. The lady who cut my hair was nice, but she wanted to leave early, so she made a mess of my head.**(Sentiment: Negative)**

5.**Original** : I fell in love with this place as soon as we pulled up and saw the lights strung up and oldies coming from the speakers! I tried the banana cream pie hard ice cream, their scoops are very generous!! My bf got the peach cobbler hard ice cream and that was to die for! We got 4 servings of ice cream for $10, which nowadays is a steal IMO! :) I'll definitely be heading back with my coworkers this week! (Sentiment: Positive)

Adversarial (TextBugger): I declined in love with this place as shortly as we pulled up and saw the headlights stung up and oldies coming from the speakers! I tried the ban ana cream pe hard ice cream, their scoops are very generous!! My bf got the peach cobbler hard ice cream and that was to die for! We got 4 servings of ice cream for $10, which nowadays is a steal IMO! :) I'll definitely be heading back with my coworkers this w eek!**(Sentiment: Negative)**

Adversarial (TextBugger+NMT-TextAttack):I fell in love with this place as soon as we pulled up and saw the lights strung up and oldies coming from the speakers! I tried the banana cream pie hard ice cream, their scoops are very generous!! My bf got the peach cobbler hard ice cream and that was to die for! We got 4 servings of ice cream for $10, which existent is a theft IMO! :) I'll doubtless be heading back with my coworkers this week! **(Sentiment: Negative)**

Back-Translated (TextBugger): I decided in love with this place as soon as we got up and climbed the chopped headlights and the old ones coming from the speakers! I've had the hard ice cream of ban ana cream, its spoonfuls are very generous! My friend got the hard iced peach pie and it was to die! We have 4 servings of ice cream for $10, which today is an OMI robbery! :) I'll definitely be coming back with my coworkers this w eek! **(Sentiment: Positive)**

Back-Translated (TextBugger+NMT-Text-Attack): I fell in love with this place as soon as we stopped and saw the stiff, old lights coming from the loudspeakers! I have tasted the hard frozen

banana cream cake, its spoonfuls are very generous!! My bf got the hard iced peach pie and he was going to die for it! We have 4 servings of ice cream for $10, which exists is an OMI robbery! :) I will definitely return with my co-workers this week!**(Sentiment: Negative)**