DOI: 10.54254/2977-3903/2/2023012

LexiGuard: Elevating NLP robustness through effortless adversarial fortification

Marwan Omar

Illinois Institute of Technology, Capitol Technology University, Chicago, USA

drmarwan.omar@gmail.com

Abstract. NLP models have demonstrated susceptibility to adversarial attacks, thereby compromising their robustness. Even slight modifications to input text possess the capacity to deceive NLP models, leading to inaccurate text classifications. In the present investigation, we introduce Lexi-Guard: an innovative method for Adversarial Text Generation. This approach facilitates the rapid and efficient generation of adversarial texts when supplied with initial input text. To illustrate, when targeting a sentiment classification model, the utilization of product categories as attributes is employed, ensuring that the sentiment of reviews remains unaltered. Empirical assessments were conducted on real-world NLP datasets to showcase the efficacy of our technique in producing adversarial texts that are both more semantically meaningful and exhibit greater diversity, surpassing the capabilities of numerous existing adversarial text generation methodologies. Furthermore, we leverage the generated adversarial instances to enhance models through adversarial training, demonstrating the heightened resilience of our generated attacks against model retraining endeavors and diverse model architectures.

Keywords: NLP models, NLP robustness, adversarial text generation

1. Introduction

Zhou et al. state that prior research has demonstrated that NLP models face difficulties due to adversarial attacks and out-of-distribution data [24]. According to Jia and Liang, Jin et al., and Alzantot et al., various research studies addressing the adversarial attack challenge have proposed the generation of adversarial examples within the input text space or intermediary representation domains [2, 14, 16]. Existing research on the construction of adversarial examples within the input text space, however, frequently demonstrates fluency issues and tends to produce adversarial instances that do not adhere to semantic restrictions and grammaticality criteria. Table 1 illustrates these limitations by highlighting the deficiencies of the available research on adversarial instances.

In this work, we aim to address limitations of prior research works in this space. Our goal is in line with Wang et al.'s exploration of the pursuit of producing adversarial cases while including controllable features [22]. We suggest utilizing text generation techniques to provide adversarial examples that are more contextually relevant and diversified. We also focus on designing perturbations that can achieve two unique goals: effectively reducing the predictive performance of an NLP model (thus resulting in inaccurate predictions) and complying to a preset set of language constraints. These goals can be achieved by using a method that is based on previous research and calls for the development of

controlled qualities in order to produce a collection of diverse, excellent adversarial examples that are semantically consistent with the original input text.

In terms of technical notation, our system refers to the input text as "x," the label for the main job (such as sentiment analysis in the context of text classification) as "y," the model's prediction for "x" as "f(x)," and the controllable attributes (such as gender, dataset domain), as "a." Our main goal is to create adversarial perturbations, marked by the letter "x," that can successfully trick the classifier into producing an inaccurate prediction, denoted by the letter "f(x)" "f(x)." We designate the change from '(x, y)' to '(x', y)' in order to assure the accuracy of our adversarial training and data annotation.

We use the Adversarial Text Generation model developed by Wang et al. to achieve this goal [22]. However, we use a completely different model architecture, apply it to a variety of datasets, and adjust the hyperparameters. The adversarial example generation model consists of an encoder and a decoder that were both trained on a large text corpus to ensure semantic coherence and adherence to linguistic restrictions. Implementing previous research's concepts and tightening the cosine similarity thresholds between sentence encodings of original and perturbed sentences and embeddings of replacement words are required to enforce semantic fidelity. By putting alterations via a grammar checker for validation, grammatical fidelity is also maintained. In addition, we incorporate grammatical and semantic limitations at each stage of the search process, emulating Moriss et al.

Real-world NLP datasets are used for our experimental validation, which demonstrates the effectiveness, adaptability, and generalizability of our suggested methodology. Our results show that the generated adversarial attacks have increased diversity, as measured by the BLEU-4 score, and have improved resistance to model retraining attempts and changes in model topologies.

2. Related work

The efficacy of employing adversarial examples as a means to enhance the robustness of Natural Language Processing (NLP) models against adversarial attacks has recently been revealed in numerous research endeavors [1, 6, 8, 12, 14, 18]. It is worth mentioning that Alzantot et al. and Jin et al. have devised methodologies for generating adversarial texts by substituting words with their synonyms, as determined by their proximity in the word embedding space [2, 16]. This approach leads to the creation of written content that perplexes models and induces erroneous classifications. Zhao et al. proposed a novel approach to creating adversarial instances in the context of continuous data representation [24]. Their research focused on exploring the semantic domain and resulted in the development of a Generative Adversarial Network (GAN) that is capable of producing coherent and comprehensible adversarial examples. Jia et al. conducted a study in a relevant context, whereby they explored word replacement combinations that aim to reduce the largest worst-case loss [15]. This investigation employed the widely utilized interval bound propagation technique. Zhu et al. have adopted a unique approach by focusing on the incorporation of adversarial perturbations into word embeddings, rather than generating textual outputs directly [25]. The objective of this approach is to mitigate adversarial risk associated with input instances.

Our research builds upon the work of Wang et al. in the field of controlled adversarial text generation [22]. Specifically, we want to enhance the performance of their model, CAT.Gen, by generating adversarial texts that are both more engaging and coherent. The scope of their suggested model's implementation is constrained to a singular dataset, specifically the Amazon Review dataset, and a singular machine learning architecture, namely the Recurrent Neural Network (RNN). However, the model demonstrates its capability to generate more authentic and contextually relevant adversarial instances in real-world scenarios. The objective of our study is to expand upon their established methodology by applying it to an alternative dataset (IMDB) and employing a transformer-based neural network. This choice is motivated by the remarkable performance exhibited by transformer models, such as BERT, in several language tasks, including sentiment categorization. The inclusion of a grammatical validation process, which ensures that the produced adversarial examples conform to proper grammar and semantic coherence, is a notable enhancement compared to the CAT-Gen model. Furthermore, a notable computational burden arises when training extensive batches of adversarial samples, which is a

DOI: 10.54254/2977-3903/2/2023012

limitation of the Wang et al. study [22]. This problem becomes particularly evident when applying their approach to a substantial dataset such as Yelp Polarity. In light of the considerable processing time we encountered, despite utilizing robust computer resources, our investigation focused on identifying a more efficient approach for generating hostile instances. To accomplish this objective, the inner ascent stages of Projected Gradient Descent (PGD) are employed. PGD is a widely used optimization technique in the field of machine learning due to its ability to efficiently compute parameter gradients without incurring significant additional computational costs when computing gradients of inputs.

The investigation we are conducting is closely connected to the topic of controllable text generation. In contrast, Iyyer et al. developed a system known as the Syntactically Controlled Paraphrase Network (SCPN) to generate adversarial instances [12]. For example, Hu et al. conducted a study exploring the utilization of variational auto-encoders in combination with holistic attribute discriminators [11]. The approach utilized in this study is founded upon an encoder-decoder model, which effectively generates adversarial training data, hence enhancing the model's resilience against adversarial attacks. Furthermore, Zhu et al. propose FreeLB, a unique technique for adversarial training that aims to enhance the level of invariance within the embedding space [25]. In this approach, the inclusion of adversarial perturbations in word embeddings is employed, and the subsequent minimization of adversarial risk is carried out across many regions including the input samples. The utilization of this approach on Transformer-based models for tasks pertaining to the comprehension of natural language and reasoning based on common sense demonstrates its effectiveness, as evidenced by the attainment of improved benchmark test outcomes for models such as BERT-base and RoBERTa-large. The authors' treatment of the temporal features of NLP models is insufficient, as they neglect to include potential temporal variations and shifts in data patterns. This issue holds significant importance, particularly in practical scenarios where the forthcoming data may exhibit variations from the characteristics and patterns observed in the training data.

3. Method

Our proposed strategy for producing adversarial cases is schematically depicted in Figure 1. Our approach is optimized for creating attacks with a narrow focus, in this case sentiment classification. This is accomplished by manipulating the property in question inside a given input sentence (in this case, product reviews). Encoder, decoder, and attribute classifier make up the model architecture; these components are also present in prior work on controlled text generation [5, 11, 21]. The system incorporates parts to make it easier to alter attributes and create assaults inside a given job model.

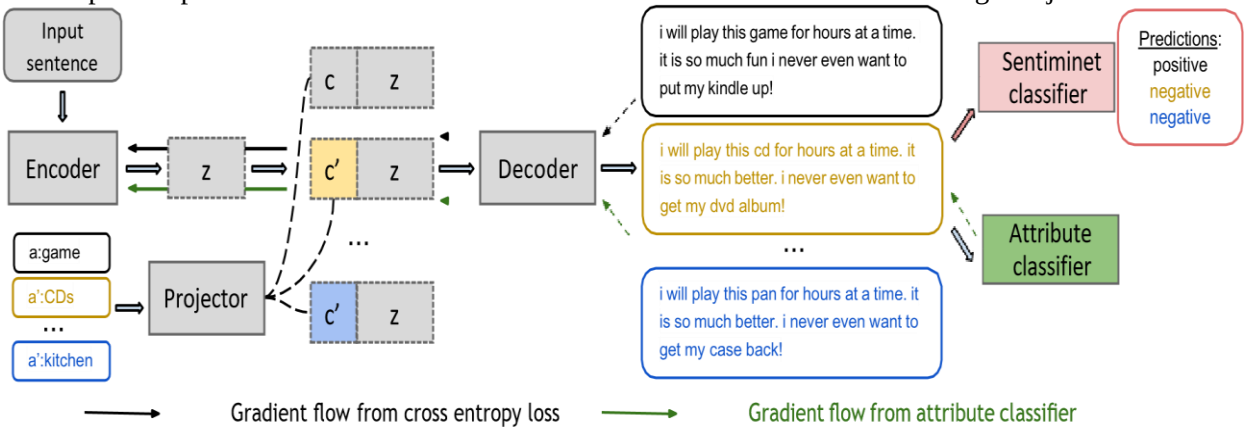

Figure 1. provides an overview of our adversarial examples generation technique.

The process of backpropagation is employed. The utilization of cross entropy loss (shown by a black dashed line) is employed to guarantee that the adversarial examples created by our system comply with grammar requirements and maintain semantic coherence. Please guarantee that the sentence generated possesses a comparable semantic meaning to the original input sentence. The manipulation of attributes (shown by the green dashed line) is used to induce attribute loss in the generated sentence, which is

DOI: 10.54254/2977-3903/2/2023012

unrelated to the task label. The prediction of sentiment labels on generated text exhibits variability while altering the attribute denoted as "a" representing different categories.

Method	Examples
Textfooler(Jin et al., 2020)	A person is relaxing on his day off → A person is relaxing on his nowadays off The two men are friends → The three men are dudes
NL-adv (Alzantot et.al., 2018)	A man is talking to his wife over his phone → A guy is chitchat to his girl over his phone A skier gets some air near a mountain… → A skier gets some airplane near a mountain…
Natural-GAN (Zhao et al., 2018)	a girl is playing at a looking man → a white preforming is lying on a beach two friends waiting for a family together → the two workers are married

Examples of adversarial text generation models that have been tested on the SNLI dataset are provided in Table 1 above [4]. Word-swapping techniques (Textfooler and NL-adv) can generate adversarial text with diverse or restricted semantics, while GAN techniques (Natural-GAN) are more likely to produce sentences that break the rules of the adversarial task.

3.1 The generation of adversarial examples at scale with high efficiency.

It is firmly believed that the optimization process is fundamental to any machine learning algorithm. Consequently, significant effort was dedicated to improving our algorithm in order to efficiently generate substantial batches of adversarial examples (AEs) and ultimately develop the most effective adversarial attacks. In order to achieve this objective, The inner ascent stages of Projected Gradient Descent (PGD), which is a well-established and highly efficient optimization technique, were utilized in our study. in the field of machine learning. By utilizing PGD, we were able to extract the gradients of the parameters with minimal computational cost while calculating the gradients of the inputs.

Algorithm 1 "Free" Large-Batch Adversarial Training (FreeLB-K)

Require: Training samples $X = \{(\boldsymbol{Z}, y)\}$, perturbation bound ϵ, learning rate τ, ascent steps K, ascent step size α

Initialize $\boldsymbol{\theta}$
for epoch $= 1 \ldots N_{ep}$ **do**
  for minibatch $B \subset X$ **do**
    $\boldsymbol{\delta}_0 \leftarrow \frac{1}{\sqrt{N_{\boldsymbol{\delta}}}} U(-\epsilon, \epsilon)$
    $\boldsymbol{g}_0 \leftarrow 0$
    for $t = 1 \ldots K$ **do**
      Accumulate gradient of parameters θ
        $\boldsymbol{g}_t \leftarrow \boldsymbol{g}_{t-1} + \frac{1}{K} \mathbb{E}_{(\boldsymbol{Z}, y) \in B} [\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X} + \boldsymbol{\delta}_{t-1}), y)]$
      Update the perturbation δ via gradient ascend
        $\boldsymbol{g}_{adv} \leftarrow \nabla_{\boldsymbol{\delta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X} + \boldsymbol{\delta}_{t-1}), y)$
        $\boldsymbol{\delta}_t \leftarrow \Pi_{\|\boldsymbol{\delta}\|_F \leq \epsilon}(\boldsymbol{\delta}_{t-1} + \alpha \cdot \boldsymbol{g}_{adv} / \|\boldsymbol{g}_{adv}\|_F)$
    end for
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \boldsymbol{g}_K$
  end for
end for

Figure 2. Optimization algorithm details to generate adversarial examples at scale

DOI: 10.54254/2977-3903/2/2023012

4. Analysis and discussion

To accomplish our goal, we employ the gong 2018 adversarial dataset sourced from IMDB. This dataset consists of binarized ratings that incorporate both positive and negative feelings. The dataset was partitioned into a training set and a test set, consisting of 25,000 reviews each. However, only 2,000 reviews from the training set were utilized for the purposes of development and testing. To allow parameter change and final evaluation, a development set and a test set are maintained, each including 10,000 occurrences. Subsequently, our classifier is trained and optimized through the application of the gradient descent optimization algorithm, utilizing both the training and development sets. The evaluation of the model's performance includes measuring its accuracy on the original instances from the test sets, as well as on adversarial examples that are generated using targeted attack methods particularly tailored for the test set. The BERT model, which is considered state-of-the-art (SOTA), will be used to classify both the attributes (category) and the task labels (sentiment) in our text. The projector employs a singular layer of Multilayer Perceptron (MLP). Throughout the progression of our research, we have seen that the training process can exhibit instability as a result of the utilization of the gumbel softmax technique for generating soft embeddings. Additionally, we have noticed that there are instances where the output sentence tends to replicate the input sentence. In accordance with the suggestions put forth by Hu et al., we meticulously calibrated the temperature parameter for the gumbel softmax [11]. Additionally, it was shown that utilizing a network with limited capacity, such as a one-layer MLP with a hidden size of 256, as the projector for the controlled attribute, along with a higher dropout ratio on sentence embeddings, such as 0.5, contributes to the stabilization of the training process. Both of these discoveries are located in the preceding section. Table 4 presents an analysis of the transferability of our instances in relation to established adversarial text generation systems, namely Jin et al. and Alzantot et al. [2,16]. The aforementioned experiments were conducted by Jin et al. and Alzantot et al.

Model Architecture	TextFooler (Jin e 2020)	NL-adv (Alzant al., 2018)	LexiGuard
Bert-retraining	84.7	82.9	48.2
WordCNN	85.6	80.5	50.6

Table 4 presents the accuracy results obtained from conducting various attacks on a re-trained model and a distinct architecture. It is important to acknowledge that the accuracy of the original model is recorded as zero due to the inclusion of a hold-out 1K set that exclusively consists of successful attacks in the evaluation process.

The results derived through qualitative analysis. Table 2 presents qualitative examples of our Lexi-Guard model. The observations indicate that the model exhibits the ability to generate adversarial texts that possess grammatical accuracy, diversity, and semantic preservation. Moreover, the model effectively substitutes many words from the initial input in order to conform to the new category attribute. The attainment of this accomplishment presents a considerable level of difficulty when relying exclusively on the substitution of synonyms or doing nearest-neighbor searches within the word embedding space. This has been exemplified in prior studies conducted by Jin et al. and Alzantot et al. [2,16]. To illustrate, our system effectively modifies the product description from "good fluffy, southern mystery" to "good fabric, not thin," thereby aligning with the attribute shift from "movie" to "shirt."

Attribute Original sentence with attribute a Generated sentence with perturbed attribute x’ ($x \rightarrow x'$)

DOI: 10.54254/2977-3903/2/2023012

Kitchen → Android	amazing knife, used for my edc for a long time, only switched because i got tired of the same old knife (Pos.)	amazing case. used for my Android for a long time, only problem because i got tired of the same old phone (Neg.)
Book → Room	not as helpful as i wanted. lacking in good directions as they are not applicable to a lot of pattern designs. (Neg.)	not as helpful as i wanted. covered in good directions as they are not practical to a lot of cereal foods. (Pos.)
Movie → Shirt	good fluffy, southern mystery. not as predictable as some. promising ending. i will probably read the rest of the series. (Pos.)	good fabric, no thin. not as predictable as pictured. last well. i will probably read the rest of the series. (Neg.)

Table 2: Successful adversarial examples generated by our Lexi-Guard model on the Movie Review Dataset.

b. Adversarial Training

The results of adversarial training, a widely used technique to enhance models by incorporating adversarial cases, are displayed in Table 3 [7]. In our study. The created adversarial samples were partitioned into two separate categories. One subset was employed to increase the size of the training data, while the other subset was used as a hold-out set for the purpose of testing. The BERT sentiment classifier model, which was previously used in Table 4, is retrained using supplemented training data. Subsequently, the model is evaluated on the hold-out set. Table 3 demonstrates the augmentation of training data with adversarial instances created by each approach, as indicated by the rows. The evaluation of model performance on the hold-out set is also conducted separately for each method, as indicated by the columns. It is evident that the incorporation of Lexi-Guard examples yields superior results in enhancing performance on Lexi-Guard assaults compared to the utilization of baselines. The baselines employ more limited substitutions, whereas the integration of Lexi-Guard examples not only enhances performance but also sustains a high level of accuracy in baseline attacks.

DOI: 10.54254/2977-3903/2/2023012

	Original test set	TextFooler attacks	NL-adv attacks	CAT-Gen attacks
Original Training	91.9	84.7	82.9	49.3
+TextFooler (Jin et al., 2020)	92.7	89.5	88.6	52.7
+NL-adv (Alzantot et al., 2018)	92.2	86.4	94.6	51.2
+Lexi-Guard	91.2	84.4	83.4	92.1

Table. In row 3, the initial training set is enhanced with the inclusion of hostile attacks. The evaluation and presentation of the accuracy of the hold-out 1K adversarial attacks, created by our approach as well as two alternative baselines, are conducted and showcased in columns.

5 Conclusion and future work

This study introduces Lexi-Guard, a paradigm that effectively generates adversarial cases while maintaining simplicity. Lexi-Guard has the capability to produce a wide range of hostile writings that are both semantically accurate and grammatically sound. Our assertion is that our model produces adversarial examples that are more pertinent to real-world issues due to the increased resilience of our attacks against model re-training and across various model architectures. The efficacy and minimal resource consumption of our technology render it attractive when compared to alternative adversarial text generators. The advantage of our approach lies specifically inside this domain. Our optimization approach additionally enables the model to acquire knowledge on the task-irrelevant attributes that are most susceptible to attack, hence enhancing its adaptability. Regarding future advancements, it would be a reasonable progression to extend the application of this methodology to diverse linguistic tasks, including but not limited to natural language inference and question answering. One such course of action that could be pursued is as follows. Additionally, it would be intriguing to assess the model's practical performance.

References

[1] David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of blackbox sequence-to-sequence models. In Pro- ceedings of the 2017 Conference on Empirical Meth- ods in Natural Language Processing, pages 412– 421, Copenhagen, Denmark. Association for Com- putational Linguistics.

[2] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang.
2018. Generating natural language adversarial ex- amples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

[3] Alexei Baevski and Michael Auli. 2018. Adaptive in- put representations for neural language modeling.

[4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large

anno- tated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

[5] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: a simple approach to controlled text generation. In ICLR.

[6] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial exam- ples for text classification. In Proceedings of the 56th Annual Meeting of the Association for Compu- tational Linguistics (Volume 2: Short Papers), pages 31–36, Melbourne, Australia. Association for Com- putational Linguistics.

[7] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In ICLR.

[8] Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. Transactions of the Association for Computational Linguistics, 6:437–450.

[9] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In Proceedings of the 25th International Conference on World Wide Web, WWW '16.

[10] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In Proceedings of the 58th Annual Meet- ing of the Association for Computational Linguis- tics.

[11] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward con- trolled generation of text. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Re- search, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

[12] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computa- tional Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

[13] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Cate- gorical reparameterization with gumbel-softmax. In ICLR.

[14] Robin Jia and Percy Liang. 2017. Adversarial exam- ples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empiri- cal Methods in Natural Language Processing, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

[15] Robin Jia, Aditi Raghunathan, Kerem Go¨ksel, and Percy Liang. 2019. Certified robustness to adver- sarial word substitutions. In Proceedings of the 2019 Conference on Empirical Methods in Natu- ral Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

[16] Di Jin, Zhijing Jin, Joey Zhou, and Peter Szolovits. 2020. Is BERT really robust? Natural language at- tack on text classification and entailment. In AAAI.

[17] Yoon Kim. 2014. Convolutional neural net- works for sentence classification. arXiv preprint arXiv:1408.5882.

[18] Aakanksha Naik, Abhilasha Ravichander, Norman M. Sadeh, Carolyn Penstein Rose´, and Graham Neubig. 2018. Stress test evaluation for natural language in- ference. In COLING.

[19] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1).

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu. 2002. BLEU: a method for automatic eval- uation of machine translation. In Proc. of ACL.

[21] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In Advances in Neural Informa- tion Processing Systems 30, pages 6830–6841.

[22] Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019. Adversar- ial domain adaptation for machine reading compre- hension. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natu- ral Language Processing (EMNLP-IJCNLP).

[23] Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In ICLR.

[24] Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analy- sis datasets: Consequences, source, and suggestions. arXiv preprint arXiv:2004.13606.

[25] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2020. Freelb: En-hanced adversarial training for language understand- ing. In ICLR.