Generating Syntactically Variant Text Adversaries

Zhengli Zhao
UC Irvine
zhengliz@uci.edu

Sameer Singh
UC Irvine
sameer@uci.edu

Abstract

Crafting adversaries for NLP models is important for exposing their vulnerabilities and evaluating their robustness. Existing approaches generate text adversaries by changing individual characters or words, or rely on back-translation models, both of which generate adversaries that consist only of short-distance, lexical edits. In this work, we propose text adversaries with substantially different syntactic structure and large edit distance from the original input while retaining its semantics. We introduce different approaches to generate such adversaries: deterministic transformations of the syntactic parses, dataset augmentation for training paraphrase generation models to produce syntactically variant examples, and models that use sentence alignment as auxiliary supervision to encourage long-distance edits. Experiments demonstrate that we can produce useful text adversaries for text classification and question answering, evaluating robustness against changes that are syntactically different.

1 Introduction

There have been several methods generating text adversaries against NLP models by inserting, modifying, or removing important words (Samanta and Mehta, 2017; Liang et al., 2018), or by injecting character-level changes (Ebrahimi et al., 2018; Belinkov and Bisk, 2018). These adversarial attack methods have been used for applications such as text classification (Lei et al., 2019) and reading comprehension (Jia and Liang, 2017), for evaluating model robustness, and even improving models via adversarial training (Niu and Bansal, 2018).

One of the key property of NLP models we are interested in studying is their generalization and robustness to the variety of syntax that can appear in real-world data. However, due to their design and the datasets they are trained on, existing methods for adversarial attacks can only generate text adversaries with short-distance and lexical edits. Generating more complex syntactic transformations, while retaining the original semantic meaning and remaining grammatical, is quite challenging. Even the model for syntactically controlled paraphrase generation in Iyyer et al. (2018) cannot generate syntactic transforms such as switching $active \leftrightarrow passive$ voice correctly, given the gold constituency parse of target sentences.

In this work, we introduce different approaches to generate text adversaries with substantially different syntactic structure and large edit distance from the input, while retaining the original semantics. We first apply a set of deterministic transformations on syntactic parses from He et al. (2015) to generate paraphrases with syntactic variants. Qualitative examples show that these syntactically different sentences can attack text classification models such as fastText (Joulin et al., 2016) for sentiment analysis. To generate syntactically different sentences that also contain lexical variations, we train a paraphrase generation model on data augmented with the syntactic transformations. We extend the training loss to encourage long-distance edits using sentence alignment annotations between paraphrases as auxiliary supervision over decoding attention. Experiments demonstrate that we produce useful text adversaries for question answering models, to evaluate their robustness against changes that are syntactically different.

2 Related Work

Adversarial examples (Goodfellow et al., 2015) are slightly perturbed instances that remain indistinguishable to humans, but change the prediction of the model. Compared to visual adversaries that add small magnitude perturbations that human cannot perceive, textual adversaries are not well-defined in the sense that human may always perceive discrete

Table 1: **Syntactic Adversaries against fastText**

rule	original vs adversarial	pred
passivization	It only recognizes the Phone as its storage device.	neg
	The Phone only is recognized by it as its storage device.	pos
quotative verbs	I just don't know how this place managed to serve the blandest food I have ever eaten when they are preparing Indian cuisine.	neg
	How this place managed to serve the blandest food I have ever eaten when they are preparing Indian cuisine, I just don't know.	pos
genitive reordering	Their rotating beer on tap is also a highlight of this place.	pos
	Their rotating beer on tap is also this place's highlight.	neg
that clause	Plus, its steep price point is worth, I seriously do not believe.	neg
	Plus, I seriously do not believe it is worth its steep price point.	pos
conjunction clause	Verizon's bills, however, are difficult to understand even though their pricing plans are simple.	neg
	Even though their pricing plans are simple, Verizon's bills, however, are difficult to understand.	pos

differences in text, and further, not all perturbations result in grammatical sentences.

Many adversarial attack methods against NLP models focus on making insertions, substitutions, and deletions at character-level (Ebrahimi et al., 2018; Belinkov and Bisk, 2018), or word-level (Samanta and Mehta, 2017; Alzantot et al., 2018; Liang et al., 2018; Li et al., 2019). They often contain heuristics of misspelling, or constraints on semantic and lexical similarity of word substitutes. Zhao et al. (2018) generate natural adversarial examples that are semantically close by using generative adversarial networks, however often change the meaning of the sentence.

Paraphrases of the original sentences that change the prediction of models have also been used as text adversaries. Existing paraphrase datasets such as ParaNMT (Wieting and Gimpel, 2018) and PAWS (Zhang et al., 2019) are collected via back-translation. These datasets contain mostly lexical variants, but minimal syntactical variants, because of the alignment between languages in back-translation. Ribeiro et al. (2018) obtain perturbations via back-translating the original sentences, and generalize them into attacking rules that are mostly local lexical replacements.

Very few have focused on generating syntactically variant text adversaries. Iyyer et al. (2018) contribute a new approach in which given a sentence and a target constituency parse, the model produces a paraphrase of the sentence with respect to the target syntax. However, their model is trained on ParaNMT-50M (Wieting and Gimpel, 2018) with mostly lexical variants, lacking of the ability to generate some syntactic transforms such as $active \leftrightarrow passive$ voice. As for adversarial attack, their model takes in predefined parse templates to generate candidates instead of finding adversarial syntax directly, which may produce ungrammatical sentences when matching incompatible templates. To improve this, we need to produce syntactic transformations given input sentences and introduce structure variants with larger edit distance into paraphrase datasets.

3 Syntax Transform Rules

As our first step to generating syntactically variant text adversaries, we adopt transformation rules that operate on constituent parse trees in He et al. (2015) to transform sentence structures while retaining semantics. The set of rules define transformations on verb phrases (passivization/activization, reordering quotation), noun phrases (*that* clause and genitive reordering), and conjunction clauses. These syntactic transformation rules are carefully designed based on linguistic analysis and can preserve grammaticality and semantic meaning. After obtaining paraphrases with syntactic transformations, we choose those that change the model predictions as adversaries.

We start with applying these rules to sentiment analysis dataset (Kotzias et al., 2015) and evalu-

ate against the fastText (Joulin et al., 2016) classification model. Picking transformed sentences that change the model predictions as adversaries, Table 1 demonstrates qualitative examples that successfully attack the fastText model, corresponding to each of these rules. With the first pair as an example, simply changing the sentence voice from *active* to *passive* can trick the model into predicting "positive" incorrectly. Compared to adversarial examples generated mainly with lexical changes, these adversaries contain more interesting long-distance edits on sentence structure that can evaluate model robustness against syntactic variants.

While these rules can produce paraphrases with syntactic variants, not all sentences can be transformed with these rules and they cannot introduce lexical changes. Instead, we are interested in training models that is capable of generating paraphrases with both lexical and syntactic variants for any input sentence.

4 Augmenting and Analyzing Datasets

To train a paraphrase model, we need the training dataset to reflect syntactic variations. We apply the above transform rules to several datasets (ParaNMT (Wieting and Gimpel, 2018), PAWS-Wiki (Zhang et al., 2019), and QQP (Iyer et al., 2017)), augmenting them with sentence pairs that are semantically similar but syntactically different.

To verify our syntactic transform augmented (STA) datasets contain more syntactically substantial variants and long-distance edits, we need to quantify the difference in word orders for a pair of sentences. To measure this distance between a pair of sentences, we first encode both source and target sentences with contextualized word embeddings BERT (Devlin et al., 2019), and then calculate the word embedding flow between source and target via Word Mover's optimization (Kusner et al., 2015). With the Word Mover Flow, we have the alignment between words and can count the *number of crossings* where aligned word pairs are crossed, similarly as defined in Zhang et al. (2019). We call the percentage of crossed word pairs over total alignment pairs the *inversion rate*: higher inversion rates imply larger long-distance changes syntactically. Figure 1 visualizes the Word Mover Flow and number of crossings for a sentence pair.

In Table 2, we compare original datasets to our augmentations using Levenshtein edit distance, number of crossings, and inversion rate. As shown,

Table 2: **Dataset Analysis.** More crossings (Xing) and higher inversion rate (inv%) imply larger word-order changes between pairs of sentences, while (edit) represents word-level Levenshtein edit distance.

tgt vs src	size	edit	Xing	inv%
ParaNMT	5m	6.29	8.44	8.3
STA-ParaNMT	1m	10.34	31.70	21.9
PAWS-Wiki	29k	5.66	19.61	6.2
STA-Wiki	28k	12.38	75.47	22.2
QQP	130k	6.09	7.12	7.2
PAWS-QQP	12k	2.46	14.39	8.7
STA-QQP	31k	9.93	24.66	21.8

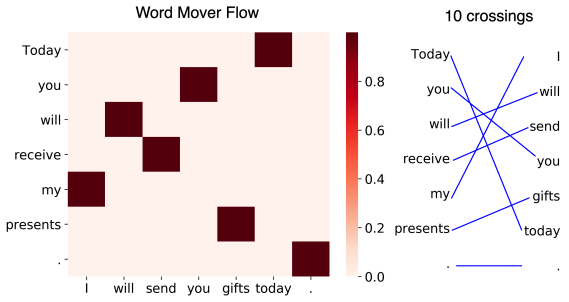

Figure 1: **Alignment from World Mover Flow**

original datasets contain much less syntactical long-distance changes than our data augmentation with transformation rules. Paraphrase datasets collected (ParaNMT) or augmented (PAWS) via back-translation introduce less syntactical variants compared to our augmentations (STA). Trained with the augmented data, models will have the potential to generate syntactically different sentences.

5 Syntactic Paraphrase Generation

To generate paraphrases with more syntactical edits and with lexical variants, we train paraphrase generation models on existing datasets and with our syntactic transform augmentations. We utilize the Pointer-Generator model (See et al., 2017) to generate paraphrases given original sentences, in which words can be copied from the source text via pointing while novel words can be produced through the generator. In addition, we use our alignment annotations between input and target sentences as auxiliary supervision over attentions for decoding.

In Table 3, we show that our model trained with augmented QQP data and alignment supervision generates sentences that consist of larger word-order changes. Table 4 reports BLEU, METEOR,

Table 3: **Ablation Study on Paraphrase Generation.** We train Pointer-Generator models with alignment supervision (+ali) on original QQP (org), or together with STA-QQP (+tf). Scores are calculated comparing generated paraphrases to source sentences.

ptr-gen vs src	edit	crossing	inv rate%
org	4.40	2.06	3.2
org +ali	4.52	2.20	3.7
org +tf	4.46	2.09	3.4
org +tf +ali	4.59	2.52	3.9

Table 4: **BLEU/METEOR/TER of Generation.** Scores are calculated comparing generated paraphrases either to source or target sentences.

ptr-gen	gen vs src	gen vs tgt
org	37.7/33.7/42.5	20.2/25.2/57.4
org +ali	40.0/34.8/40.7	21.3/25.7/56.9
org +tf	40.5/35.0/40.2	20.6/25.4/57.0
org +tf +ali	39.1/34.4/41.4	21.7/25.9/56.1

and TER scores of generated sentences compared to the input and target sentences, confirming that alignments lead to higher quality sentences.

On May 21, 2013, NFL owners at their spring meetings in Boston voted and awarded the game to Levi's Stadium. The $1.2 billion stadium opened in 2014. It is the first Super Bowl held in the San Francisco Bay Area since Super Bowl XIX in 1985, and the first in California since Super Bowl XXXVII took place in San Diego in 2003.
org: When was the last time California hosted a Super Bowl? 1985
adv: When was the last time a Super Bowl was hosted by California ? 2003

6 Syntactically Variant Text Adversaries

We use our paraphrase generation model trained on QQP with augmented STA-QQP to rewrite questions in the SQuAD dataset (Rajpurkar et al., 2016), and attack the questions for pretrained BiDAF model (Seo et al., 2017) with context paragraph unchanged. We present examples of different syntactical edits with original questions and corresponding predicted answers in blue and adversarial ones in red. More examples with lexical and syntactical changes are available in Appendix A. These adversarial examples contain syntactical edits that cannot be generated by existing attack methods, and are useful in evaluating models against syntactic variants. Our paraphrase model also learns lexical variants and does not require constituent parse trees to produce syntactic transformations.

Newton came to realize that the effects of gravity might be observed in different ways at larger distances. In particular, Newton determined that the acceleration of the Moon around the Earth could be ascribed to the same force of gravity if the acceleration due to gravity decreased as an inverse square law.
org: How might gravity effects be observed differently according to Newton? in different ways at larger distances
adv: According to Newton, how might one observe gravity effects differently? inverse square law

CBS broadcast Super Bowl 50 in the U.S., and charged an average of $5 million for a 30-second commercial during the game. The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyonc and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively. It was the third-most watched U.S. broadcast ever.
org: What was the average cost of a 30-second commercial? $5 million
adv: What was a 30-second commercial's average cost? 30-second

7 Discussion

In this work, we focus on generating syntactically variant text adversaries, and demonstrate interesting adversarial examples in text classification and question answering. We augment paraphrase datasets with syntactic transformation rules and improve model with alignment supervision to generate syntactically variant text adversaries. In future work, we can train models on different domains of data (e.g. Wiki) and apply adversarial attack against other NLP applications such as machine translation and dialogue conversation.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

He He, Alvin Grissom II, John Morgan, Jordan Boyd-Graber, and Hal Daumé III. 2015. Syntax-based rewriting for simultaneous machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. *https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning (ICML)*, pages 957–966.

Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G Dimakis, Inderjit S Dhillon, and Michael Witbrock. 2019. Discrete attacks and submodular optimization with applications to text classification. In *Conference on Systems and Machine Learning (SysML)*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *Network and Distributed System Security Symposium (NDSS)*.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. In *Conference on Computational Natural Language Learning (CoNLL)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations (ICLR)*.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*.

A Adversarial Examples against BiDAF on SQuAD

Following the CretaceousPaleogene extinction event, the extinction of the dinosaurs and the wetter climate may have allowed the tropical rainforest to spread out across the continent. From 6634 Mya, the rainforest extended as far south as 45. Climate fluctuations during the last 34 million years have allowed savanna regions to expand into the tropics. During the Oligocene, for example, the rainforest spanned a relatively narrow band. It expanded again during the Middle Miocene, then retracted to a mostly inland formation at the last glacial maximum. However, the rainforest still managed to thrive during these glacial periods, allowing for the survival and evolution of a broad diversity of species.
org: Did the rainforest managed to thrive during the glacial periods? the rainforest still managed to thrive
adv: During the glacial periods, did the rainforest managed to thrive? glacial periods

The success of any pathogen depends on its ability to elude host immune responses. Therefore, pathogens evolved several methods that allow them to successfully infect a host, while evading detection or destruction by the immune system. Bacteria often overcome physical barriers by secreting enzymes that digest the barrier, for example, by using a type II secretion system. Alternatively, using a type III secretion system, they may insert a hollow tube into the host cell, providing a direct route for proteins to move from the pathogen to the host. These proteins are often used to shut down host defenses.
org: In a type III secretion system, proteins are transported to the host cell in order to do what? move from the pathogen to the host
adv: Proteins are transported to the host cell in order to do what in a type III secretion system? insert a hollow tube

Non-revolutionary civil disobedience is a simple disobedience of laws on the grounds that they are judged "wrong" by an individual conscience, or as part of an effort to render certain laws ineffective, to cause their repeal, or to exert pressure to get one's political wishes on some other issue. Revolutionary civil disobedience is more of an active attempt to overthrow a government (or to change cultural traditions, social customs, religious beliefs, etc...revolution doesn't have to be political, i.e. "cultural revolution", it simply implies sweeping and widespread change to a section of the social fabric). Gandhi's acts have been described as revolutionary civil disobedience. It has been claimed that the Hungarians under Ferenc Dek directed revolutionary civil disobedience against the Austrian government. Thoreau also wrote of civil disobedience accomplishing "peaceable revolution." Howard Zinn, Harvey Wheeler, and others have identified the right espoused in The Declaration of Independence to "alter or abolish" an unjust government to be a principle of civil disobedience.
org: What is the goal of individual civil disobedience? render certain laws ineffective
adv: What is individual civil disobedience's goal? peaceable revolution

Gasquet (1908) claimed that the Latin name atra mors (Black Death) for the 14th-century epidemic first appeared in modern times in 1631 in a book on Danish history by J.I. Pontanus: "Vulgo & ab effectu atram mortem vocatibant". ("Commonly and from its effects, they called it the black death"). The name spread through Scandinavia and then Germany, gradually becoming attached to the mid 14th-century epidemic as a proper name. In England, it was not until 1823 that the medieval epidemic was first called the Black Death.
org: Who claimed that the name Black Death first appeared in 1631? J.I. Pontanus
adv: Who claimed that in 1631, the name Black Death first appeared? Gasquet