

Improving Robustness by Augmenting Training Sentences with Predicate-Argument Structures

Nafise Sadat Moosavi[‡], Marcel de Boer[†], Prasetya Ajie Utama^{‡†}, Iryna Gurevych[†]

[‡]Ubiquitous Knowledge Processing Lab (UKP-TUDA)

[†]Research Training Group AIPHES

Department of Computer Science, Technische Universität Darmstadt

<https://www.ukp.tu-darmstadt.de>

Abstract

Existing NLP datasets contain various biases, and models tend to quickly learn those biases, which in turn limits their robustness. Existing approaches to improve robustness against dataset biases mostly focus on changing the training objective so that models learn less from biased examples. Besides, they mostly focus on addressing a specific bias, and while they improve the performance on adversarial evaluation sets of the targeted bias, they may bias the model in other ways, and therefore, hurt the overall robustness. In this paper, we propose to augment the input sentences in the training data with their corresponding predicate-argument structures, which provide a higher-level abstraction over different realizations of the same meaning and help the model to recognize important parts of sentences. We show that without targeting a specific bias, our sentence augmentation improves the robustness of transformer models against multiple biases. In addition, we show that models can still be vulnerable to the lexical overlap bias, even when the training data does not contain this bias, and that the sentence augmentation also improves the robustness in this scenario. We will release our adversarial datasets to evaluate bias in such a scenario as well as our augmentation scripts at <https://github.com/UKPLab/data-augmentation-for-robustness>.

1 Introduction

Due to annotation artifacts, existing datasets contain certain biases.¹ Models often rely on these biases to perform well on the corresponding evaluation set, which also includes similar biases (Gururangan et al., 2018; Poliak et al., 2018a; McCoy

et al., 2019b; Gardner et al., 2020). As a result, the model learns the spurious patterns in the data instead of the intended phenomena of the dataset, which in turn limits the robustness and makes models vulnerable against adversarial evaluations (McCoy et al., 2019b; Nie et al., 2019a). The adversarial evaluation sets consist of counterexamples in which relying on the bias results in incorrect predictions. Overcoming such biases is an important challenge in developing robust NLP models.

The majority of existing works improve the robustness against a given bias by proposing new methods or training paradigms (He et al., 2019; Clark et al., 2019; Mahabadi and Henderson, 2019; Utama et al., 2020a,b; Wu et al., 2020). The common component in such methods is a bias model that is trained to detect training examples that can be solved only using a bias. This information is then used for ignoring or down-weighting biased training examples (He et al., 2019; Clark et al., 2019; Mahabadi and Henderson, 2019), or tuning the confidence of the model on such examples (Utama et al., 2020a). While these methods are very effective in improving the robustness against the targeted bias, they have two shortcomings:

- They mostly target a specific bias and discourage the model from learning that bias. However, they may bias the model in other unwanted directions and therefore hurt the overall robustness.²
- They are only applicable to scenarios in which the training examples contain the bias. However, as we show in this paper, a model can still be vulnerable to a specific bias even if

¹In this work, the term bias refers to the label bias as defined by Shah et al. (2020), i.e., the conditional distribution of the target label diverges at test time based on specific attributes of the training data.

²The concurrent work of Utama et al. (2020b) and Wu et al. (2020) are the exceptions to this trend, in which they address multiple biases together and show that their debiasing methods improve the overall robustness.

the training examples do not explicitly exhibit that bias.

An alternative approach is to augment the training data with additional counterexamples for the bias (McCoy et al., 2019b; Elkahky et al., 2018). This may also result in overfitting to the augmented counterexamples and hurting the overall robustness (Nie et al., 2019a).

In this paper, we propose to augment existing training sentences with their corresponding predicate-argument structures. The motivation of using predicate-argument structures is to provide a higher-level abstraction over different surface realizations of the same underlying meaning. We examine the impact of this linguistic augmentation on pre-trained transformers, e.g., BERT (Devlin et al., 2019a), which achieve state-of-the-art performances on numerous NLP datasets. The addition of predicate-argument structures to input sentences helps the model to recognize and focus on more important parts of sentences, and therefore, to learn a different attention pattern.

The findings of this paper are as follows:

- We show that the model may still be vulnerable to a specific bias even when training examples of the target task do not contain that bias. We propose new adversarial sets for evaluating the robustness of models in such scenarios based on the SWAG dataset and the lexical overlap bias. Lexical overlap is a common bias in various NLP datasets, e.g., natural language inference (McCoy et al., 2019b), or question answering (Jia and Liang, 2017a). We show that the performance of pre-trained transformers that are fine-tuned on the SWAG dataset (Zellers et al., 2018) drop below a random baseline on evaluation sets that contain this bias.
- Our results show that without targeting a specific bias or adding additional training examples, the proposed sentence augmentation improves the robustness of the model against various types of biases. Besides, we show that the sentence augmentation is effective in both scenarios when the training data contains or does not contain the bias. Our approach only requires the augmentation for the training sentences and does not require any changes to the test data, and therefore, does not add any additional cost at the test time.

- Our results emphasize the importance of evaluating the impact of debiasing methods on more than one adversarial set and with more than one base model.

2 Related Work

Debiasing Methods. Existing debiasing solutions fall into one of these two categories: (1) extending the training data with additional counterexamples, and (2) proposing a new approach that recognizes biased examples in the training data and then using this knowledge during training (Utama et al., 2020a; He et al., 2019; Clark et al., 2019; Mahabadi and Henderson, 2019; Utama et al., 2020b; Wu et al., 2020).

The first type of solution is to identify the bias and augment the training data with counterexamples in which relying on the targeted bias results in incorrect predictions. While augmenting the training data with counterexamples improves the results on the targeted bias, it may hurt the overall robustness (Nie et al., 2019a). As mentioned by Jha et al. (2020), while augmentation with counterexamples helps the model to unlearn the targeted bias, it is unlikely that it encourages the model to rely on more generalizable features of the data.

The approaches of the second category first use a bias detection model for recognizing training examples that contain the bias. They then either (a) train an ensemble of the bias model and a base model so that the base model only learns from non-biased examples (He et al., 2019; Clark et al., 2019; Mahabadi and Henderson, 2019), (b) change the importance of the biased training examples in the training objective (Schuster et al., 2019; Mahabadi and Henderson, 2019), or (c) change the confidence of the model on biased examples (Utama et al., 2020a).

The shortcoming of existing debiasing methods is that they mostly model a single bias and only evaluate the impact of the proposed method on the adversarial evaluation set of the targeted bias. Therefore, while they improve the performance on the targeted adversarial sets, they may hurt the overall robustness. The recent work of Utama et al. (2020b) and Wu et al. (2020) are the exceptions in which they show that their proposed debiasing frameworks improve the overall robustness, and hence the generalization across different datasets in natural language understanding and question answering, respectively. Utama et al. (2020b) propose

a new framework that automatically recognizes biased training examples and does not require pre-defining bias types. Therefore, the recognized biased examples may contain various bias types. [Wu et al. \(2020\)](#) propose a framework for modeling multiple known biases concurrently. To do so, they propose to combine two bias weights in the training objective including (a) a dataset-level weight indicating the strength of the bias in the datasets, and (b) an example-level weight indicating the strength of the bias in a training example. The common finding in both works is that debiasing based on multiple biases is a key factor in improving the overall robustness.

Compared to existing debiasing methods:

- Our proposed approach does not include an additional model to recognize biased examples or additional training examples. As we show in Section 4.4, since it does not target any specific bias, it improves the robustness of the baseline model against multiple biases.
- Since it does not require recognizing biased examples, it is applicable to improve robustness against biases that do not exist in the training examples.

Using Linguistic Structures for Neural Models.

The use of linguistic information in recent neural models is not very common. The use of such information has been mainly investigated for tasks in which there is a clear relation between the linguistic features and the target task. For instance, various neural models use syntactic information for the task of semantic role labeling (SRL) ([Roth and Lapata, 2016](#); [Marcheggiani and Titov, 2017](#); [Strubell et al., 2018](#); [Swayamdipta et al., 2018](#)), which is closely related to syntactic relations, i.e., some arcs in the syntactic dependency tree can be mirrored in semantic dependency relations.

[Marcheggiani and Titov \(2017\)](#) build a graph representation from the input text using their corresponding dependency relations and use graph convolutional networks (GCNs) to process the resulting graph for SRL. They show that the incorporation of syntactic relations improves the in-domain but decreases the out-of-domain performance.

Similarly, [Cao et al. \(2019\)](#) and [Dhingra et al. \(2018\)](#) incorporate linguistic information, i.e., coreference relations, in their model and show improvements in in-domain evaluations.

[Strubell et al. \(2018\)](#) use linguistic information, i.e., dependency parse, part-of-speech tags, and predicates for SRL using a transformer-based encoder ([Vaswani et al., 2017](#)). They make use of this linguistic information by (1) using multi-task learning, and (2) supervising the neural attention of the transformer model to predict syntactic dependencies. They use gold syntax information during training and predicted information during the test time. Their model substantially improves both in-domain and out-of-domain performance in SRL. However, these results are then outperformed by a simple BERT model without using any additional linguistic information ([Shi and Lin, 2019](#)).

[Moosavi and Strube \(2018\)](#) examine the use of various linguistic features, e.g., syntactic dependency relations and gender and number information, as additional input features to a neural coreference resolver. They show that using informative linguistic features substantially improves the generalization of the examined model.

In a similar direction, [Moosavi et al. \(2019\)](#) improve the robustness by enhancing the input representations. They did so by adding a set of simple features to the input where the input is a pair of text sequences and show that it improves generalization across similar datasets and tasks.

All the above approaches require additional linguistic information, e.g., syntax, both during the training and the test time. [Swayamdipta et al. \(2018\)](#), on the other hand, only make use of the additional syntactic information during training. They use multi-task learning by considering syntax parsing as an auxiliary task and minimizing the combination of the losses of the main and auxiliary tasks. They use syntactic information for the tasks of SRL and coreference resolution. They show that this information slightly improves the in-domain performance. In this work, we do not change the loss function and only augment the input sentences of the training data. The advantage of our solution is that it does not require any changes in the model or its training objective. It can be applied to all the transformer-based models without changing the training procedure.

Using Predicate-Argument Structures.

Predicate-argument structures have been used for improving the performance of downstream tasks like machine translation ([Liu and Gildea, 2010](#); [Bazrafshan and Gildea, 2013](#)), reading comprehension ([Berant et al., 2014](#); [Wang et al.,](#)

2015), and dialogue systems (Tur et al., 2005; Chen et al., 2013). However, these approaches are based on pre-neural models.

The proposed model by Marcheggiani et al. (2018) for neural machine translation is a sample neural model that incorporates predicate-argument structures. Unlike this work, Marcheggiani et al. (2018) incorporate these linguistic structures at the model-level. They add two layers of semantic GCNs on top of a standard encoder, e.g., convolutional neural network or bidirectional LSTM. The semantic structures are used for determining nodes and edges in the GCNs. In this work, however, we incorporate these structures at the input level, and only for the training data. Therefore, we can use the state-of-the-art models without any changes.

Overall, this work differs from the related work because (1) it evaluates the use of predicate-argument structures for improving the robustness of transformer-based models on natural language understanding tasks, and (2) it uses these structures at the input level to extend raw inputs, (3) it only employs this information during training, and (4) it requires no changes in the model or the training procedure.

3 Augmenting Input Sentences with Predicate-argument Structures

We augment the raw text of each input sentence in the training data with its corresponding predicate-argument structures. We use the PropBank-style semantic role labeling model of Shi and Lin (2019), which has state-of-the-art results on the CoNLL-2009 dataset. We specify the beginning of the augmentation by the [PRD] special token that indicates that the next tokens are the detected predicate.³ We then specify the ARG0 and ARG1 arguments, if any, with [AG0] and [AG1] special tokens, respectively. The end of the detected predicate-argument structure is also specified by the [PRE] special token. If more than one predicate is detected for a sentence, we at most add the first three detected predicate-argument structures.⁴ Figure 1 shows an example for an augmented sentence.

³Special tokens are atomic, i.e., they are not split by the tokenizer.

⁴In our preliminary experiments, we found out that this setting works better than adding all of them.

<p>Original: Someone takes the drink, then holds it. Augmented: Someone takes the drink, then holds it. [PRD] takes [AG0] Someone [AG1] the drink [PRE] [PRD] holds [AG0] Someone [AG1] it [PRE]</p>

Figure 1: Augmenting the text of an input sentence with its predicate-argument structures.

4 Impact of Sentence Augmentation on Improving Robustness

In this section, we explain the datasets and models that we use in our experiments, as well as the result of the sentence augmentation compared to other recent debiasing methods.

4.1 Training Data

As mentioned, we have evaluated the impact of the proposed augmentation on two different settings: (1) when the training data contains the investigated biases, and (2) when training examples do not explicitly contain the bias.

Training Data with Biases (MultiNLI). For evaluating the impact of data augmentation when the training examples are biased, we use MultiNLI (Williams et al., 2018) for training, which is the largest available dataset for Natural Language Inference (NLI). Given a premise and a hypothesis, NLI is the task of determining whether the hypothesis is entailed, contradicts, or is neutral to the premise.

Various studies show that similar to many other NLP datasets, MultiNLI contains various biases. For instance, hypothesis sentences may contain words that are highly associated with a target label, regardless of the premise (Gururangan et al., 2018; Poliak et al., 2018b). This bias is referred to as the hypothesis-only bias. Another well-known bias in MultiNLI is the lexical overlap bias, i.e., the label of most premise-hypothesis pairs with overlapping words is entailment.

Training Data without the Bias (SWAG). For the second setting, we evaluate the impact of the augmentation to improve the robustness against the lexical overlap bias. Lexical overlap is a common bias in various NLP datasets, e.g., NLI (McCoy et al., 2019b), duplicate question detection (Zhang et al., 2019), or question answering (Jia and Liang, 2017b). We use the SWAG dataset (Zellers et al., 2018) as the training data.

Given a premise about a situation, the task of the SWAG dataset, i.e., grounded commonsense reasoning, is to reason about what is happening and to predict what might come next. The task is modeled as a multiple choice answer selection. For instance, “The tutorial starts by showing each part of the drum set up close” is a correct ending for the premise “A man in a black polo shirt is sitting in front of an electronic drum set”.

If we train the bias model of Clark et al. (2019) for solving the task only based on lexical overlap features, it only achieves 26% accuracy on SWAG, which is around the random baseline, while it achieves 65% accuracy on MultiNLI.⁵ This indicates that the examples in the SWAG dataset are not affected by this bias. Please note that the SWAG dataset may also contain various biases. However, it does not contain the lexical overlap bias.

4.2 Evaluation Sets

In this section, we discuss the adversarial evaluation sets that we use to evaluate the robustness of the model. Apart from the adversarial sets, which are the **out-of-distribution** evaluation sets compared to the training data distribution, we also report the performance on the corresponding development set as the **in-domain** performance.

4.2.1 MultiNLI Evaluation Sets

We use the following adversarial sets for evaluating models that are trained on MultiNLI:

MultiNLI Hard: Gururangan et al. (2018) introduce a hard split for MultiNLI evaluation sets in which models cannot predict the correct labels using the hypothesis-only bias.

HANS: McCoy et al. (2019b) create this dataset for evaluating the lexical overlap bias. Sentence pairs in HANS include various forms of lexical overlap, namely *lexical overlap*, *subsequence*, and *constituent*.

In the *lexical overlap* subset, all words of the hypothesis appear in the premise. For instance, “The doctor was paid by the actor” and “The doctor paid the actor” sentence pair belong to this subset.

The *subsequence* subset contains hypotheses, which are a contiguous subsequence of their corresponding premise. “The doctor near the actor danced” and “The actor danced” are sample sentence pairs from this subset.

⁵The details of this bias model is reported in the supplementary materials.

Finally, in the *constituent* subset, hypotheses are a complete subtree of the premise. For example, “If the artist slept, the actor ran” and “The artist slept” belong to this subset.

Stress Test: Naik et al. (2018) provide adversarial evaluation sets based on weaknesses of state-of-the-art NLI models. We use *negation*, *word overlap*, and *length mismatch* sets from the stress test, in which a tautology is added at the end of the premise or hypothesis in MultiNLI.

- **Negation:** the tautology “and false is not true” is added to the end of all the hypothesis sentences in the MultiNLI development set for creating this evaluation set. The presence of the negation word “not” may confuse the model to predict contradiction.
- **Word Overlap:** For creating this evaluation set, Naik et al. (2018) append the tautology “and true is true” to the end of all the hypothesis sentences in the MultiNLI development set.
- **Length Mismatch:** the tautology “and true is true” is appended five times to the end of the premise sentences in the MultiNLI development set for creating this adversarial evaluation set.

4.2.2 SWAG Evaluation Sets

We created three different adversarial datasets based on the SWAG development set for evaluating the lexical overlap bias. These datasets evaluate the model’s understanding of (1) syntactic variations, (2) antonym relations, and (3) named entities in the presence of high lexical overlap.

The common property in all three evaluation sets is a high lexical overlap between the sentence pairs. In all these evaluation sets, one of the incorrect endings is replaced with a new incorrect ending that has a high lexical overlap with the premise. Since the new incorrect endings are created automatically, they may contain sentences that are not meaningful. For instance, the syntactic variations subset contains the incorrect endings “a key holds up someone” and “The last page flips to the writer” for the “Someone holds up a key” and “The writer flips to the last page” premises, respectively. Humans can recognize such sentences are not meaningful, and therefore, they are not a plausible ending for given premises. However, as we will see, because of the

lexical overlap bias, the model mostly selects the new incorrect endings.

Syntactic Variations: In this evaluation set, we take premises that contain subject-verb-object structures from the SWAG development set. We then construct a new negative ending by swapping the subject and object of the premise and replace one of the existing negative endings with the new one.⁶ This dataset includes 20K samples. It is similar to a subset of the *lexical overlap* subset in HANS, as well as the adversarial evaluation that is explored by Nie et al. (2019b) for NLI.

Antonym Relations: In this test set, we create a new negative ending by replacing the first verb of the premise with its antonym. We use WordNet for antonym relations. This adversarial setting is also common in NLI, e.g., (Naik et al., 2018; Gloeckner et al., 2018). As an example, “A lot of people are *standing* on terraces in a big field and people is walking in the entrance of a big stadium” is an incorrect ending for the “A lot of people are *sitting* on terraces in a big field and people is walking in the entrance of a big stadium” premise in this evaluation set. This set contains 7476 samples.

Named Entities: In this adversarial dataset, a new incorrect ending is created by replacing one of the named entities of the premise with an unrelated named entity.⁷ For instance, based on the “The reflection he sees is *Harrison Ford* as someone Solo winking back at him” premise, we create “The reflection he sees is *Eve* as someone Solo winking back at him.” as the new incorrect ending. This test set contains 190 samples.

4.3 Base Model

We use the Bert-base-uncased model (Devlin et al., 2019b) as the base model.⁸ *Bert-orig* refers to the results when the model is trained on the original training data. *Bert-aug* represents the results when the base model is trained on the augmented data. The set of all parameters are the same in *Bert-orig* and *Bert-aug*. Besides, the evaluation data is the same for both *Bert-orig* and *Bert-aug* experiments, and their only difference is their training data.

We compare our results with the confidence regularization approach of Utama et al. (2020a) and the

product-of-expert approach (He et al., 2019; Clark et al., 2019). They both use Bert-base-uncased as the base model.

CR(lex) and *CR(hypo)* refer to the confidence-regularization method when it is debiased based on the lexical overlap and hypothesis-only biases, respectively. Similarly, *POE(lex)* and *POE(hypo)* show the product-of-expert results based on the lexical overlap and hypothesis-only biases, respectively.

We use the same set of hyper-parameters⁹ for all the models and report the average performance using five different random seeds for all the results, i.e., Table 1-Table 4. As reported by McCoy et al. (2019a) and Zhou et al. (2020), the performance on the adversarial evaluation sets can highly vary given different values of hyper-parameters.¹⁰ Therefore, to ensure a fair comparison, results on adversarial sets should be reported using the same parameters in out-of-distribution evaluations.

4.4 Results

Table 1 and Table 2 show the results of the examined models on all evaluation sets of MultiNLI and SWAG datasets, respectively. MultiNLI and its corresponding adversarial evaluation sets, i.e., HARD and Stress Test, contain two subsets, matched and mismatched. Sentence pairs in the matched subsets are from the same domain as those of the training data while they are from different domains in the mismatched subsets. We have reported the results on the matched subsets in Table 1. The results on the mismatched subsets are included in the supplementary material, and they follow the same pattern.

The confidence regularization and product-of-expert debiasing methods model a single bias at a time and train a bias detection model to detect training examples that can be solved by only using biased features. Therefore, they are only applicable when training examples contain the examined bias, and they are not used for the experiments of Table 2.

Based on the results of Table 1:

- The model that is debiased for a specific bias has a higher accuracy on the corresponding adversarial evaluation set, i.e., *POE(lex)* has the highest average score on HANS. However,

⁶We use Stanford parser (Chen and Manning, 2014) for detecting subjects and objects.

⁷We use the Stanford named entity recognizer (Finkel et al., 2005) for determining the named entities.

⁸We use Huggingface Transformers (Wolf et al., 2019).

⁹I.e., batch size=16, learning rate=2e-5, and the same set of random seeds.

¹⁰E.g., The accuracy on the non-entailment examples in the lexical overlap subset of HANS can vary between 6% to 54% for the BERT-base model using different random seeds.

Model	In-domain		HANS		Hard	Negation	Stress Test	
		lex.	subs.	const.			Overlap	Length
BERT-orig	84.2±0.3	62.9±7.8	52.1±0.9	56.0±1.4	75.3±0.6	55.5±0.6	59.5±1.1	81.3±0.3
CR(lex)	83.7±0.1	62.1±2.7	60.9±3.7	64.2±1.6	74.8±0.3	55.6±0.4	59.5±1.0	81.3±0.3
CR(hypo)	84.4±0.2	68.5±8.3	53.9±1.2	56.2±0.8	77.2±0.4	55.1±0.4	59.5±1.3	81.5±0.3
PoE(lex)	82.6±0.2	69.1±5.9	68.1±9.6	68.8±3.6	73.2±0.2	55.6±0.4	59.1±0.8	80.7±0.4
PoE(hypo)	83.3±0.4	66.6±7.3	53.1±1.2	58.8±2.3	77.8±0.7	55.2±0.5	59.7±0.8	80.5±0.3
BERT-aug	84.4±0.1	69.8±5.6	53.5±1.5	58.3±2.0	75.9±0.3	56.3±0.2	60.2±1.4	81.5±0.3

Table 1: Comparing the impact of the augmentation to the confidence regularization (CR) (Utama et al., 2020a), and product-of-expert (POE) (He et al., 2019; Clark et al., 2019) methods debiased for the lexical overlap (lex) and hypothesis-only (hypo) biases. E.g., CR(lex) is the confidence regularization method debiased based on lexical overlap. All models are trained on MultiNLI with the same hyperparameters. Highest scored on each dataset are boldfaced. Scores that are lower than BERT-orig are marked in gray.

Model	Dev.	Syntax	Antonym	NEs
BERT-base	81.1±0.1	27.7±0.9	18.3±1.7	7.9±1.1
BERT-aug	79.1±0.3	47.1±1.4	36.3±1.6	15.9±1.7

Table 2: Results on SWAG and its adversarial sets.

they can reduce the performance on the non-targeted evaluation sets, e.g., *POE(lex)* results are below the baseline on HARD and Stress Test evaluation sets.

- The use of sentence augmentation results in consistent improvements across the examined evaluation sets.

Based on the results of Table 2, we see that while BERT has very high accuracy on the SWAG development set, its performance drops below a random baseline on the lexical overlap adversarial sets. This indicates that the model is biased towards selecting the endings which have a high lexical overlap to the premise, while the training data does not contain this bias. Besides, we see that augmenting the training data improves the accuracy on the adversarial sets from 8-20 points.

An example of the attention pattern with and without sentence augmentation Figure 2 shows the difference of the BERT attention weights, using BertViz¹¹ (Vig, 2019), on an example from the HANS dataset. In this example, the premise and hypothesis are “The senators supported the secretary in front of the doctor.” and “The doctor supported the senators.”, respectively.

For instance, for the predicate “supported” in the hypothesis, the BERT model that is trained on *augmented* data (bottom subfigure), has high attention weights on “senators”, “supported”, and

“secretary”, while for *original* the attention weights of this predicate are more distributed. Similarly, for the subject “doctor” in the hypothesis, *augmented* mainly attends to the corresponding subject in the premise, i.e., “senators”.

5 Are the improvements model-agnostic?

We also evaluate the impact of our augmentation on other transformer models including XLNET (Yang et al., 2019) and RoBERTa (Liu et al., 2019).

The differences of the examined transformer models are as follows: BERT is jointly trained on a masked language modeling task and a next sentence prediction task. It is pre-trained on the Book-Corpus and English Wikipedia. XLNET (Yang et al., 2019) is trained with a permutation-based language modeling objective for capturing bidirectional contexts. The XLNet-base model is trained with the same data as BERT-base. The RoBERTa model (Liu et al., 2019) has the same architecture as BERT. However, it is trained with dynamic masking and without the next sentence prediction task. It is also trained using larger batch-size, vocabulary size, and training data.

Table 3 and Table 4 present the results when the models are trained on the original vs. augmented MultiNLI and SWAG training data, respectively.

The results show that when the model itself is relatively robust and has high performance on various evaluation sets, which is the case for RoBERTa trained on MultiNLI, augmenting the training data will not have a positive impact. While several recent debiasing methods are model-agnostic, e.g., (Utama et al., 2020a; Mahabadi and Henderson, 2019; Clark et al., 2019), they are evaluated using only the BERT base model. This result indicates the importance of evaluating model-agnostic meth-

¹¹<https://github.com/jessevig/bertviz>

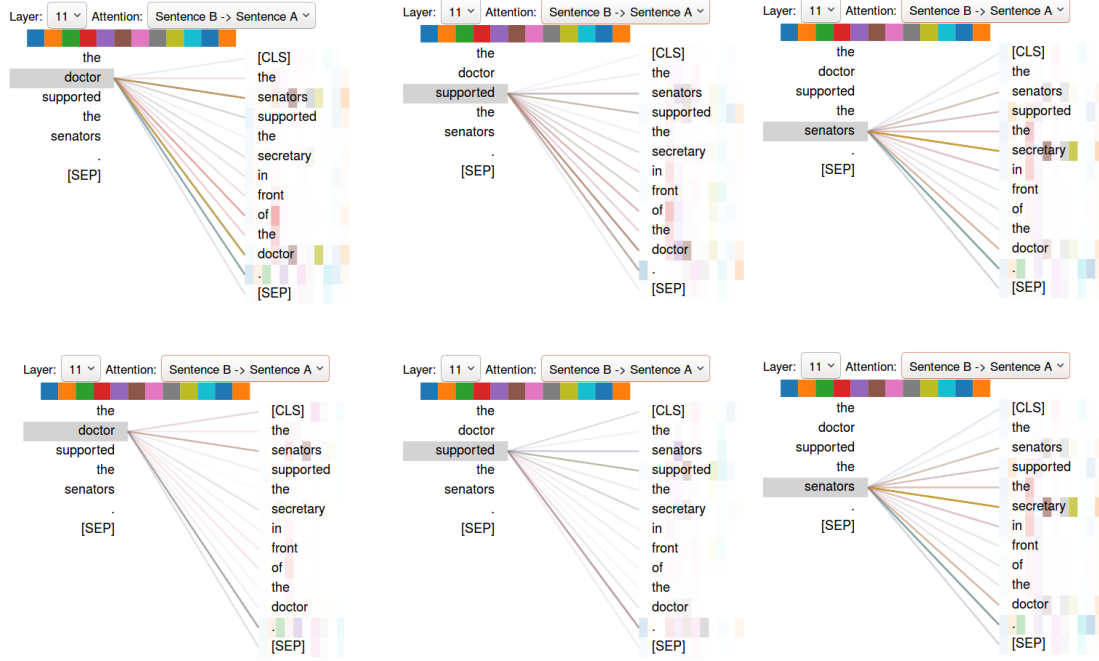


Figure 2: BERT attention weights on an example from the HANS dataset based on *original* (top weights) and *augmented* (bottom weights) training. Attention weights are visualized using BertViz (Fig, 2019). They highlight the attention between the hypothesis and premise words and for the predicate-argument structures of the hypothesis.

Model	In-domain		HANS		Hard		Stress Test		SNLI	
		lex.	subs.	const.		Negation	Overlap	Length		
XLNET-orig	86.6 \pm 0.1	70.2 \pm 1.3	53.6 \pm 0.9	66.5 \pm 2.2	79.4 \pm 0.8	57.6 \pm 1.2	62.7 \pm 0.7	84.0 \pm 0.4	80.9 \pm 0.1	
XLNET-aug	86.1 \pm 0.3	71.5 \pm 1.1	55.4 \pm 1.2	66.6 \pm 4.7	79.0 \pm 0.6	60.6 \pm 1.5	67.5 \pm 4.1	84.1 \pm 0.1	80.7 \pm 0.3	
RoBERTa-orig	87.5 \pm 0.1	83.3 \pm 2.3	64.6 \pm 1.0	68.1 \pm 3.3	80.6 \pm 0.1	57.3 \pm 0.4	65.5 \pm 1.8	85.0 \pm 0.2	84.0 \pm 0.1	
RoBERTa-aug	87.3 \pm 0.2	82.1 \pm 1.7	61.8 \pm 0.8	66.3 \pm 3.5	80.3 \pm 0.2	57.2 \pm 0.5	63.7 \pm 1.2	85.0 \pm 0.1	83.8 \pm 0.4	

Table 3: Impact of linguistic augmentation on the XLNET and RoBERTa models trained on MultiNLI. Accuracy scores that are higher than the baseline are boldfaced.

Model	Dev.	Syntax	Antonym	NEs
XLNET-orig	80.2 \pm 0.2	37.0 \pm 1.6	25.4 \pm 1.7	17.1 \pm 2.8
XLNET-aug	77.8 \pm 0.3	63.3 \pm 1.1	50.5 \pm 1.2	36.7 \pm 3.8
RoBERTa-orig	83.7 \pm 0.2	49.1 \pm 2.4	31.7 \pm 2.0	24.1 \pm 1.3
RoBERTa-aug	82.0 \pm 0.3	60.9 \pm 1.1	47.7 \pm 1.3	38.6 \pm 4.0

Table 4: Impact of sentence augmentation on XLNET and RoBERTa trained on SWAG.

ods on more than one model.

6 Conclusions

We propose a new approach for improving the robustness of transformer models by augmenting the training sentences with their corresponding predicate-argument structures. We show that without targeting any specific bias, sentence augmentation improves the robustness against different types of biases. Sentence augmentation is independent of the underlying task and model and therefore applies

to different tasks and settings. The augmentation only applies to the training examples, and therefore, it does not add any additional complexity at the test time. We evaluate the impact of the proposed augmentation on the natural language inference and grounded common sense reasoning tasks. However, this work opens new research directions on improving robustness by using better linguistically-informed input representations, rather than simply using raw texts. To ensure improved robustness, we encourage the community to evaluate their debiasing methods on (1) more than one evaluation set, (2) in a wider setting in which the bias does or does not exist in the training data, and (3) with more than one base model.

Acknowledgments

We would like to thank Kevin Stowe, Jan-Christoph Klie, Zahra Ahmadi, and the anonymous reviewers for their constructive feedbacks. This work is funded by the German Research Foundation through the research training group AIPHEs (GRK 1994/1) and by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

- Marzieh Bazrafshan and Daniel Gildea. 2013. [Semantic roles for string to tree machine translation](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–423, Sofia, Bulgaria. Association for Computational Linguistics.
- Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. [Modeling biological processes for reading comprehension](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.
- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. [Question answering by reasoning across documents with graph convolutional networks](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, page to appear. Association for Computational Linguistics.
- Danqi Chen and Christopher Manning. 2014. [A fast and accurate dependency parser using neural networks](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.
- Yun-Nung Chen, William Yang Wang, and Alex Rudnicky. 2013. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *Proceedings of the ASRU 2013*, Olomouc, Czech Republic. IEEE.
- Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. [Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4067–4080, Hong Kong, China. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2018. [Neural models for reasoning over multiple mentions using coreference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 42–48, New Orleans, Louisiana. Association for Computational Linguistics.
- Ali Elkahky, Kellie Webster, Daniel Andor, and Emily Pitler. 2018. [A challenge set and methods for noun-verb ambiguity](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2572, Brussels, Belgium. Association for Computational Linguistics.
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. [Incorporating non-local information into information extraction systems by Gibbs sampling](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.
- Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. [Evaluating nlp models via contrast sets](#). *arXiv preprint arXiv:2004.02709*.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that require simple lexical inferences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

- He He, Sheng Zha, and Haohan Wang. 2019. [Unlearn dataset bias in natural language inference by fitting the residual](#). In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.
- Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. [When does data augmentation help generalization in nlp?](#) *arXiv preprint arXiv:2004.15012*.
- Robin Jia and Percy Liang. 2017a. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Robin Jia and Percy Liang. 2017b. [Adversarial examples for evaluating reading comprehension systems](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- Ding Liu and Daniel Gildea. 2010. [Semantic role features for machine translation](#). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 716–724, Beijing, China. Coling 2010 Organizing Committee.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Rabeeh Karimi Mahabadi and James Henderson. 2019. [simple but effective techniques to reduce biases](#). *arXiv preprint arXiv:1909.06321*.
- Diego Marcheggiani, Joost Bastings, and Ivan Titov. 2018. [Exploiting semantics in neural machine translation with graph convolutional networks](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 486–492, New Orleans, Louisiana. Association for Computational Linguistics.
- Diego Marcheggiani and Ivan Titov. 2017. [Encoding sentences with graph convolutional networks for semantic role labeling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1506–1515, Copenhagen, Denmark. Association for Computational Linguistics.
- R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2019a. [Berts of a feather do not generalize together: Large variability ingeneralization across models with similar test set performance](#). Technical report, Department of Cognitive Science, Johns Hopkins University.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019b. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Nafise Sadat Moosavi and Michael Strube. 2018. [Using linguistic features to improve the generalization capability of neural coreference resolvers](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Brussels, Belgium. Association for Computational Linguistics.
- Nafise Sadat Moosavi, Prasetya Ajie Utama, Andreas Rücklé, and Iryna Gurevych. 2019. [Improving generalization by incorporating coverage in natural language inference](#). *arXiv preprint arXiv:1909.08940*.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. [Stress test evaluation for natural language inference](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019a. [Analyzing compositionality-sensitivity of nli models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6867–6874.
- Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019b. [Analyzing compositionality-sensitivity of nli models](#). In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018a. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191. Association for Computational Linguistics.
- Michael Roth and Mirella Lapata. 2016. [Neural semantic role labeling with dependency path embeddings](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202, Berlin, Germany. Association for Computational Linguistics.
- Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. [Towards debiasing fact verification](#)

- models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.
- Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. [Predictive biases in natural language processing models: A conceptual framework and overview](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.
- Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.
- Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. [Linguistically-informed self-attention for semantic role labeling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.
- Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. [Syntactic scaffolds for semantic structures](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.
- Gokhan Tur, Dilek Hakkani-Tur, and Ananlada Chotimongkol. 2005. [Semi-supervised learning for spoken language understanding semantic role labeling](#). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 232 – 237.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. [Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance](#). *arXiv preprint arXiv:2005.00315*, page to appear.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. *arXiv preprint arXiv:2009.12303*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Jesse Vig. 2019. [Visualizing attention in transformer-based language representation models](#). *arXiv preprint arXiv:1904.02679*.
- Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. [Machine comprehension with syntax, frames, and semantics](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 700–706, Beijing, China. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R’emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Mingzhu Wu, Nafise Sadat Moosavi, Andreas Rckl, and Iryna Gurevych. 2020. [Improving QA generalization by concurrent modeling of multiple biases](#). In *Proceedings of the Findings of ACL: EMNLP 2020*, Online. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. [SWAG: A large-scale adversarial dataset for grounded commonsense inference](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.
- Yuan Zhang, Jason Baldridge, and Luheng He. 2019. [PAWS: Paraphrase adversaries from word scrambling](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.
- Xiang Zhou, Yixin Nie, Hao Tan, and Mohit Bansal. 2020. The curse of performance instability in analysis datasets: Consequences, source, and suggestions. *arXiv preprint arXiv:2004.13606*.

A Lexical Overlap Bias Model

Clark et al. (2019) propose a simple model that uses a non-linear classifier on top of a set of simple

Model	In-domain	Hard	Stress Test		
			Negation	Overlap	Length
BERT-base	84.7 \pm 0.2	77.1 \pm 0.2	56.1 \pm 0.7	59.1 \pm 0.6	82.3 \pm 0.3
CR (lex)	84.2 \pm 0.2	76.2 \pm 0.3	56.2 \pm 0.5	58.9 \pm 0.8	82.3 \pm 0.2
CR (hypo)	84.8 \pm 0.2	78.8 \pm 0.5	55.7 \pm 0.4	59.4 \pm 1.0	82.4 \pm 0.3
PoE (lex)	82.9 \pm 0.1	74.3 \pm 0.2	55.8 \pm 0.4	58.4 \pm 0.6	81.7 \pm 0.1
PoE (hypo)	83.7 \pm 0.4	79.1 \pm 0.6	55.6 \pm 0.5	59.3 \pm 0.9	81.4 \pm 0.5
BERT-aug	84.7 \pm 0.1	77.3 \pm 0.2	56.5 \pm 0.3	59.4 \pm 0.6	82.6 \pm 0.2

Table 5: Comparing the impact of the augmentation to the confidence regularization (CR) (Utama et al., 2020a), and product-of-expert (POE) (He et al., 2019; Clark et al., 2019) methods debiased for the lexical overlap (lex) and hypothesis-only (hypo) biases. E.g., CR(lex) is the confidence regularization. All the results are reported on the mismatched subset of each evaluation set.

features including: (1) whether all the hypothesis words exist in the premise, (2) whether the hypothesis is a subsequence of the premise, (3) the fraction of hypothesis words that also exist in the premise, and (4) the max and the average of the cosine distances between the premise and hypothesis word vectors. This model is used for detecting training examples that can be solved using the lexical overlap bias.

B Results on the mismatched evaluation sets

Table 5 presents the corresponding results of Table 1 on the mismatched subsets.