Measure and Improve Robustness in NLP Models: A Survey

Xuezhi Wang
Google Research
xuezhiw@google.com

Haohan Wang
Carnegie Mellon University
haohanw@cs.cmu.edu

Diyi Yang
Georgia Institute of Technology
dyang888@gatech.edu

Abstract

As NLP models achieved state-of-the-art performances over benchmarks and gained wide applications, it has been increasingly important to ensure the safe deployment of these models in the real world, e.g., making sure the models are robust against unseen or challenging scenarios. Despite robustness being an increasingly studied topic, it has been separately explored in applications like vision and NLP, with various definitions, evaluation and mitigation strategies in multiple lines of research. In this paper, we aim to provide a unifying survey of how to define, measure and improve robustness in NLP. We first connect multiple definitions of robustness, then unify various lines of work on identifying robustness failures and evaluating models' robustness. Correspondingly, we present mitigation strategies that are data-driven, model-driven, and inductive-prior-based, with a more systematic view of how to effectively improve robustness in NLP models. Finally, we conclude by outlining open challenges and future directions to motivate further research in this area.

1 Introduction

NLP models, especially with the recent advances of large pre-trained language models have achieved great progress and gained wide applications in the real world. Despite the performance gains, NLP models are still fragile and brittle to out-of-domain data (Hendrycks et al., 2020a; Wang et al., 2019d), adversarial attacks (McCoy et al., 2019; Jia and Liang, 2017; Jin et al., 2020), or small perturbation to the input (Ebrahimi et al., 2018; Belinkov and Bisk, 2018). Those failures could hinder the safe deployment of these models in the real world, and impact NLP models' trustworthiness to users. As a result, an increasing line of work has been conducted to understand robustness issues in the language technologies communities. Still, diverse sets of research across multiple dimensions and

numerous levels of depth exist and are scattered across various communities; for instance, using a variety of definitions on a wide range of very different NLP tasks. In this work, we provide a unifying overview of what is robustness in NLP, how to identify robustness failures and evaluate model's robustness, and systematic ways to improve robustness, as well as a conceptual schema categorizing ongoing research directions. We identify gaps between the to-date robustness work, the technical opportunities, and discuss possible paths forward.

2 Definitions of Robustness in NLP

Robustness, despite its specific definitions in various lines of research, can typically be unified as follows: denote the input as x, and its associated gold label for the main task as y, assume a model f is trained on $(x, y) \sim \mathcal{D}$ and its prediction over x as $f(x)$; now given test data $(x', y') \sim \mathcal{D}' \neq \mathcal{D}$, we can measure a model's robustness by its performance on $\mathcal{D}'$, e.g., using the model's robust accuracy (Tsipras et al., 2019; Yang et al., 2020), defined as $\mathbb{E}_{(x',y')\sim D'}[f(x') = y']$. Existing literature on robustness in NLP can be roughly categorized by how $\mathcal{D}'$ is constructed: by synthetically perturbing the input (Section 2.1), or $\mathcal{D}'$ is naturally occurring with a distribution shift (Section 2.2).

The above definition works for a range of NLP tasks like text classification and sequence labeling where y is defined over a fixed set of discrete labels. For tasks like text generation, robustness is less well defined and can manifest as positional bias (Jung et al., 2019; Kryscinski et al., 2019), or hallucination (Maynez et al., 2020; Parikh et al., 2020; Zhou et al., 2021). One major challenge here is a lack of robust metrics in evaluating the quality of the generated text (Sellam et al., 2020; Zhang et al., 2020b), i.e., we need a reliable metric to determine the relationship between $f(x')$ and y' when both are open-ended texts.

2.1 Robustness against Adversarial Attacks

In one line of research, $\mathcal{D}'$ is constructed by perturbations around input x to form x' (x' typically being defined within some proximity of x). This topic has been widely explored in computer vision under the concept of adversarial robustness, which measures models' performances against carefully crafted noises generated deliberately to deceive the model to predict wrongly, pioneered by (Szegedy et al., 2013; Goodfellow et al., 2015), and later extended to NLP, such as (Ebrahimi et al., 2018; Alzantot et al., 2018; Li et al., 2019; Feng et al., 2018; Kuleshov et al., 2018; Jia et al., 2019; Zang et al., 2020; Pruthi et al., 2019; Wang et al., 2019e; Garg and Ramakrishnan, 2020; Tan et al., 2020a,b; Schwinn et al., 2021; Li et al., 2021; Boucher et al., 2022) and multilingual adversaries (Yang et al., 2019; Tan and Joty, 2021). The generation of adversarial examples primarily builds upon the observation that we can generate samples that are meaningful to humans (e.g., by perturbing the samples with changes that are imperceptible to humans) while altering the prediction of the models for this sample. In this regard, human's remarkable ability in understanding a large set of synonyms (Li et al., 2020) or interesting characteristics in ignoring the exact order of letters (Wang et al., 2020b) are often opportunities to create adversarial examples. A related line of work such as data-poisoning (Wallace et al., 2021) and weight-poisoning (Kurita et al., 2020) exposes NLP models' vulnerability against attacks during the training process. One can refer to more comprehensive reviews and broader discussions on this topic in Zhang et al. (2020c) and Morris et al. (2020b).

Assumptions around Label-preserving and Semantic-preserving Most existing work in vision makes a relatively simplified assumption that the gold label of x' remains unchanged under a bounded perturbation over x, i.e., $y' = y$, and a model's robust behaviour should be $f(x') = y$ (Szegedy et al., 2013; Goodfellow et al., 2015). A similar line of work in NLP follows the same label-preserving assumption with small text perturbations like token and character swapping (Alzantot et al., 2018; Jin et al., 2020; Ren et al., 2019; Ebrahimi et al., 2018), paraphrasing (Iyyer et al., 2018; Gan and Ng, 2019), semantically equivalent adversarial rules (Ribeiro et al., 2018), and adding distractors (Jia and Liang, 2017). However, this label-preserving assumption might not always hold, e.g., Wang et al. (2021b) studied several existing text perturbation techniques and found that a significant portion of perturbed examples are **not** label-preserving (despite their label-preserving assumptions), or the resulting labels have a high disagreement among human raters (i.e., can even fool humans). Morris et al. (2020a) also call for more attention to the *validity* of perturbed examples for a more accurate robustness evaluation.

Another line of work aims to perturb the input x to x' in small but meaningful ways that explicitly *change* the gold label, i.e., $y' \neq y$, under which case the robust behaviour of a model should be $f(x') = y'$ and $f(x') \neq y$ (Gardner et al., 2020; Kaushik et al., 2019; Schlegel et al., 2021). We believe these two lines of work are complementary to each other, and both should be explored in future research to measure models' robustness more comprehensively.

One alternative notion is whether the perturbation from x to x' is "semantic-preseving" (Alzantot et al., 2018; Jin et al., 2020; Ren et al., 2019) or "semantic-modifying" (Shi and Huang, 2020; Jia and Liang, 2017). Note this is slightly different from the above label-preserving assumptions, as it is defined over the perturbations on (x, x') rather than making an assumption on (y, y'), e.g., semantic-modifying perturbations can be either label-preserving (Jia and Liang, 2017; Shi and Huang, 2020) or label-changing (Gardner et al., 2020; Kaushik et al., 2019).

2.2 Robustness under Distribution Shift

Another line of research focuses on (x', y') drawn from a different distribution that is naturally-occurring (Hendrycks et al., 2021), where robustness can be defined around model's performance under distribution shift. Different from work on domain adaptation (Patel et al., 2015; Wilson and Cook, 2020) and transfer learning (Pan and Yang, 2010), existing definitions of robustness are closer to the concept of domain generalization (Muandet et al., 2013; Gulrajani and Lopez-Paz, 2021), or out-of-distribution generalization to unforeseen distribution shifts (Hendrycks et al., 2020a), where the test data (either labeled or unlabeled) is assumed not available during training, i.e., generalization without adaptation. In the context of NLP, robustness to natural distribution shifts can also mean models' performance should not degrade due to the differences in grammar errors, dialects, speakers,

languages (Craig and Washington, 2002; Blodgett et al., 2016; Demszky et al., 2021), or newly collected datasets for the same task but in different domains (Miller et al., 2020). Another closely connected line of research is fairness, which has been studied in various NLP applications, see (Sun et al., 2019) for a more in-depth survey in this area. For example, gendered stereotypes or biases have been observed in NLP tasks including co-reference resolution (Zhao et al., 2018a; Rudinger et al., 2017), occupation classification (De-Arteaga et al., 2019), and neural machine translation (Prates et al., 2019; Font and Costa-jussà, 2019).

2.3 Connections and A Common Theme

The above two categories of robustness can be unified under the same framework, i.e., whether $\mathcal{D}'$ represents a *synthetic* distribution shift (via adversarial attacks) or a *natural* distribution shift. Existing work has shown a model's performance might degrade substantially in both cases, but the *transferability* of the two categories is relatively underexplored. In the vision domain, Taori et al. (2020) investigate models' robustness to natural distribution shift, and show that robustness to synthetic distribution shift might offer little to no robustness improvement under natural distribution shift. Some studies show NLP models might not generalize to *unseen* adversarial patterns (Huang et al., 2020; Jha et al., 2020; Joshi and He, 2021), but more work is needed to systematically bridge the gap between NLP models' robustness under natural and synthetic distribution shifts.

To better understand *why* models exhibit a lack of robustness, some existing work attributed this to the fact that models sometimes utilize *spurious* correlations between input features and labels, rather than the *genuine* ones, where *spurious* features are commonly defined as features that do not causally affect a task's label (Srivastava et al., 2020; Wang and Culotta, 2020b): they correlate with task labels but fail to transfer to more challenging test conditions or out-of-distribution data (Geirhos et al., 2020). Some other work defined it as "prediction rules that work for the majority examples but do not hold in general" (Tu et al., 2020). Such spurious correlations are sometimes referred as dataset bias (Clark et al., 2019; He et al., 2019), annotation artifacts (Gururangan et al., 2018), or group shift (Oren et al., 2019) in the literature. Further, evidence showed that controlling model's learning

in spurious features will improve model's performances in distribution shifts (Wang et al., 2019a,b); also, discussions on the connections between adversarial robustness and learning of spurious features have been raised (Ilyas et al., 2019; Wang et al., 2020a). Theoretical discussions connecting these fields have also been offered by crediting a reason of model's lack of robustness in either distribution shift or adversarial attack to model's learning of spurious features (Wang et al., 2021c).

Further, in certain applications, model "robustness" can also be connected with models' instability (Milani Fard et al., 2016), or models having poorly-calibrated uncertainty estimation (Guo et al., 2017), where Bayesian methods (Graves, 2011; Blundell et al., 2015), dropout-based (Gal and Ghahramani, 2016; Kingma et al., 2015) and ensemble-based approaches (Lakshminarayanan et al., 2017) have been proposed to improve models' uncertainty estimation. Recently, Ovadia et al. (2019) have shown models' uncertainty estimation can degrade significantly under distributional shift, and call for more work to ensure a model "knows when it doesn't know" by giving lower uncertainty estimates over out-of-distribution data. This is another example where models can be less robust under distributional shifts, and again emphasizes the need of building more unified benchmarks to measure a model's performance (e.g., robust accuracy, calibration, stability) under distribution shifts, in addition to in-distribution accuracy.

3 Robustness in Vision vs. in NLP

Despite the widely study of robustness in vision, the study of robustness in NLP cannot always directly borrow the ideas. We categorize the main differences with the three following points:

Continuous vs. Discrete in Search Space The most obvious characteristic is probably the discrete nature of the space of text. This particularly posed a challenge towards the adversarial attack and defense regime when the study in vision is transferred to NLP (Lei et al., 2019; Zhang et al., 2020c), in the sense that simple gradient-based adversarial attacks will not directly translate to meaningful attacks in the discrete text space, and multiple novel attack methods are proposed to fill the gap, as we will discuss in later sections.

Perceptible to Human vs. Not On a related topic, one of the most impressive property of ad-

versarial attack in vision is that small perturbation of the image data imperceptible to human are sufficient to deceive the model (Szegedy et al., 2013), while this can hardly be true for NLP attacks. Instead of being imperceptible, the adversarial attacks in NLP typically are bounded by the fact that the meaning of the sentences are not altered (despite being perceptible). On the other hand, there are ways to generate samples where the changes, although being perceptible, are often ignored by human brain due to some psychological prior on how a human processes the text (Anastasopoulos et al., 2019; Wang et al., 2020b).

Support vs. Density Difference of the Data Distributions Another difference is more likely seen in the discussion of the domain adaptation of vision and NLP study. In vision study, although the images from training distribution and test distribution can be sufficiently different, the train and test distributions mostly share the same support (the pixels are always sampled from a 0-255 integer space), although the density of these distributions can be very different (e.g., photos vs. sketches). On the other hand, domain adaptation of NLP sometimes studies the regime where the supports of the data differ, e.g., the vocabularies can be significantly different in cross-lingual studies (Abad et al., 2020; Zhang et al., 2020a).

A Common Theme Despite the disparities between vision and NLP, the common theme of pushing the model to generalize from $\mathcal{D}$ to $\mathcal{D}'$ preserves. The practical difference between $\mathcal{D}$ and $\mathcal{D}'$ is more than often defined by the human's understanding of the data, and can differ in vision and NLP as humans perceive and process images and texts in subtly different ways, which creates both opportunities for learning and barriers for direct transfer. Certain lines of research try to bridge the learning in the vision domain to the embedding space in the NLP domain, while other lines of research create more interpretable attacks in the discrete text space (see Table 1 for these two lines of work). How those two lines of research transfer to each other, or complement each other, is not fully explored and calls for additional research.

4   Identify Robustness Failures

As robustness gained increasing attention in NLP literature, various lines of work have proposed ways to identify robustness failures in NLP models.

Existing works can be roughly categorized by *how* the failures are identified, among which a large portion of work relies on human priors and error analyses over existing NLP models (Section 4.1), and other lines of work adopt model-based approaches (Section 4.2). The identified robustness failure patterns are usually organized into challenging/adversarial benchmark datasets to more accurately measure an NLP model's robustness. In Table 1, we organize commonly used perturbation types for identifying models' robustness failures, and in Table 2 we summarize common robustness benchmarks for each NLP task.

4.1   Human Prior and Error Analyses Driven

An increasing body of work has been conducted on understanding and measuring robustness in NLP models (Tu et al., 2020; Sagawa et al., 2020b; Geirhos et al., 2020) across various NLP tasks, largely relying on human priors and error analyses.

Natural Language Inference Naik et al. (2018) sampled misclassified examples and analyzed their potential sources of errors, which are then grouped into a typology of common reasons for error. Such error types then served as the bases to construct the *stress test* set, to further evaluate whether NLI models have the ability to make real inferential decisions, or simply rely on sophisticated pattern matching. Gururangan et al. (2018) found that current NLI models are likely to identify the label by relying only on the hypothesis, and Poliak et al. (2018) provided similar augments that using a hypothesis-only model can outperform a set of strong baselines. Kaushik et al. (2019) asked humans to generate counterfactual NLI examples, to better understand what features are causal and encourage models to learn those features.

Question Answering Jia and Liang (2017) proposed to generate adversarial QA examples by concatenating an adversarial distracting sentence at the end of a paragraph. Miller et al. (2020) built four new test sets for the Stanford Question Answering Dataset (SQuAD) and found most question-answering systems fail to generalize to this new data, calling for new evaluation metrics towards natural distribution shifts.

Machine Translation Belinkov and Bisk (2018) found that character-based neural machine translation (NMT) models are brittle under noisy data, where noises (e.g., typos, misspellings, etc) are

Space	Perturbation level	Methods
Discrete	Character-level	HotFlip (Ebrahimi et al., 2018), DeepWordBug (Gao et al., 2018), Synthetic-Noise (Karpukhin et al., 2019)
	Word-level	GenAdv (Alzantot et al., 2018), PWWS (Ren et al., 2019), SEM (Wang et al., 2019e), BERT-ATTACK (Li et al., 2020), TextFooler (Jin et al., 2020), SememePSO (Zang et al., 2020)
	Sentence-level	AdvSQuAD (Jia and Liang, 2017), SCPNs (Iyyer et al., 2018), CAT-Gen (Wang et al., 2020c), TAILOR (Ross et al., 2021)
	Mixed-types	CheckList (Ribeiro et al., 2020), Polyjuice (Wu et al., 2021), MAYA (Chen et al., 2021c)
Continuous	Embedding space	AT & VAT (Miyato et al., 2017), Natural-adversary (Zhao et al., 2018b), FreeLB (Zhu et al., 2020), ALUM (Liu et al., 2020)

Table 1: Perturbation types for identifying robustness failures and improving robustness in NLP.

Task	Robustness Benchmarks
Natural Language Inference	Stress-test (Naik et al., 2018), HANS (McCoy et al., 2019), Counterfactual-NLI (Kaushik et al., 2019), ANLI (Nie et al., 2020)
Question Answering	AdvSQuAD (Jia and Liang, 2017), Adv-QA (Bartolo et al., 2020), Natural-Perturbed-QA (Khashabi et al., 2020), Natural-shift-QA (Miller et al., 2020), SAM (Schlegel et al., 2021)
Paraphrase Identification	PAWS (Zhang et al., 2019b), PAWS-X (Yang et al., 2019), Modify-with-Shared-Words (Shi and Huang, 2020)
Co-reference	WinoGender (Rudinger et al., 2018), WinoBias (Zhao et al., 2018a)
Named Entity Recognition	OntoRock (Lin et al., 2021), SeqAttack (Simoncini and Spanakis, 2021)

Table 2: A list of robustness benchmarks (challenging or adversarial datasets) and their corresponding tasks.

synthetically generated using possible lexical replacements. Data augmentation with artificially-introduced grammatical errors (Anastasopoulos et al., 2019) or with random synthetic noises (Vaibhav et al., 2019; Karpukhin et al., 2019) can make the system more robust to such spurious patterns. On the other hand, Wang et al. (2020b) showed another approach by limiting the input space of the characters so that the models will be likely to perceive data typos and misspellings.

Syntactic and Semantic Parsing Robust parsing has been studied in several existing works (Lee et al., 1995; Aït-Mokhtar et al., 2002). More recent work showed that neural semantic parsers are still not robust against lexical and stylistic variations, or meaning-preserving perturbations (Marzinotto et al., 2019; Huang et al., 2021), and proposed ways to improve their robustness through data augmentation (Huang et al., 2021) and adversarial learning (Marzinotto et al., 2019).

Text Generation Existing work found that text generation models also suffer from robustness issues, e.g., text summarization models suffer from positional bias (Jung et al., 2019), layout bias

(Kryscinski et al., 2019), and a lack of faithfulness and factuality (Kryscinski et al., 2019; Maynez et al., 2020; Chen et al., 2021b); data-to-text models sometimes hallucinate texts that are not supported by the data (Parikh et al., 2020; Wang et al., 2020d). In addition, Sellam et al. (2020); Zhang et al. (2020b) pointed out the deficiency of existing automatic evaluation metrics and proposed new metrics to better align the generation quality with human judgements.

Connection with Dataset Biases The robustness failures can sometimes be attributed to dataset biases, i.e., biases introduced during dataset collection (Fouhey et al., 2018) or human annotation artifacts (Gururangan et al., 2018; Geva et al., 2019; Rudinger et al., 2017), which could affect how well a model trained from this dataset generalizes, and how accurately we estimate a model's performance. For example, Lewis et al. (2021) show there is a significant test-train data overlap in a set of open-domain question-answering benchmarks, and many QA models perform substantially worse on questions that cannot be memorized from training data. In natural language inference, McCoy et al. (2019)

show that commonly used crowdsourced datasets for training NLI models might make certain syntactic heuristics more easily adopted by statistical learners. Further, Bras et al. (2020) propose to use a lightweight adversarial filtering approach to filter dataset biases, which is approximated using each instance's predictability score.

4.2 Model-based Identification

In addition to the human-prior and error-analysis driven approaches which are usually specific to each task, other lines of work identify robustness failures that are *task-agnostic* like white-box text attack methods (Ebrahimi et al., 2018; Alzantot et al., 2018; Jin et al., 2020), and even *input-agnostic* like universal adversarial triggers (Wallace et al., 2019a) and natural attack triggers (Song et al., 2021).

Another line of work proposes to learn an additional model to capture biases, e.g., in visual question answering, Clark et al. (2019) train a naive model to predict prototypical answers based on the question only irrespective of the context; He et al. (2019); Utama et al. (2020a) propose to learn a biased model that only uses dataset-bias related features. This framework has also been used to capture unknown biases assuming that the lower capacity model learns to capture relatively shallow correlations during training (Clark et al., 2020). In addition, Wang and Culotta (2020a) identify model shortcuts by training classifiers to better distinguish "spurious" correlations from "genuine" ones based on human annotated examples.

Model-in-the-loop vs. Human-in-the-loop Some work adopts human-in-the-loop to generate challenging examples, e.g., Counterfacutal-NLI (Kaushik et al., 2019) and Natural-Perturbed-QA (Khashabi et al., 2020). Other work applies model-in-the-loop to increase the likelihood that the perturbed examples are challenging for state-of-the-art models, but it might also introduce biases towards the particular model used. For example, SWAG (Zellers et al., 2018) was introduced that fooled most models at the time of publishing but was soon "solved" after BERT (Devlin et al., 2019) was introduced. As a result, Yuan et al. (2021) present a study over the transferability of adversarial examples, and Contrast Sets (Gardner et al., 2020) intentionally avoid using model-in-the-loop. Further, more recent work adopts adversarial human-and-model-in-the-loop to create more difficult examples for benchmarking, e.g., Adv-QA (Bartolo et al., 2020), Adv-Quizbowl (Wallace et al., 2019b), ANLI (Nie et al., 2020), and Dynabench (Kiela et al., 2021).

5 Improve Model Robustness

Correspondingly, there are multiple lines of directions that try to improve robustness in NLP models. Depending on where and how the intervention is applied, those approaches can be categorized into the following categories: data-driven (Section 5.1), model-based and training-scheme-based (Section 5.2), inductive-prior-based (Section 5.3) and finally causal intervention (Section 5.4).

5.1 Data-driven Approaches

Data augmentation recently gained a lot of interest, in improving performance in low-resourced language settings, few-shot learning, mitigating biases, and improving robustness in NLP models (Feng et al., 2021; Dhole et al., 2021). Techniques like Mixup (Zhang et al., 2018), MixText (Chen et al., 2020), CutOut (DeVries and Taylor, 2017), AugMix (Hendrycks et al., 2020b), HiddenCut (Chen et al., 2021a), have been shown to substantially improve the robustness and the generalization of models. Such mitigation strategies are operated at the data level, and often hard to be interpreted in terms of how and why mitigation works.

Other lines of work deal with spans or regions associated within data points to prevent models from heavily relying on spurious patterns. To make NLP models more robust on sentiment analysis and NLI tasks, Kaushik et al. (2019) proposed curating counterfactually augmented data via a human-in-the-loop process, and showed that models trained on the combination of this augmented data and original data are less sensitive to spurious patterns. Differently, Wang et al. (2021d) performed strategic data augmentation to perturb the set of "short-cuts" that are automatically identified, and found that mitigating these leads to more robust models in multiple NLP tasks. This line of mitigation strategies closely relates to how spurious correlations can be measured and identified, as many of the challenging or adversarial examples (Table 1) can sometimes be used to augment the original model to improve its robustness, either in the discrete input space as additional training examples (Liu et al., 2019; Kaushik et al., 2019; Anastasopoulos et al., 2019; Vaibhav et al., 2019; Khashabi et al., 2020), or in the embedding space (Zhu et al., 2020; Zhao

et al., 2018b; Miyato et al., 2017; Liu et al., 2020).

5.2 Model and Training-based Approaches

Pre-training   Recent work has demonstrated pre-training as an effective way to improve NLP models' out-of-distribution robustness (Hendrycks et al., 2020a; Tu et al., 2020), potentially due to its self-supervised objective and the use of large amounts of diverse pre-training data that encourages generalization from a small number of examples that counter the spurious correlations. Tu et al. (2020) showed a few other factors can also contribute to robust accuracy, including larger model size, more fine-tuning data, and longer fine-tuning. A similar observation is made by Taori et al. (2020) in the vision domain, where the authors found training with larger and more diverse datasets offer better robustness consistently in multiple cases, compared to various robustness interventions proposed in the existing literature.

Training with a Better Use of Minority Examples   Further, there are several works that propose to robustify the models via a better use of minority examples, e.g., examples that are under-represented in the training distribution, or examples that are harder to learn. For example, Yaghoobzadeh et al. (2021) proposed to first fine-tune the model on the full data, and then on minority examples only.

In general, the training strategy with an emphasis on a subset of samples that are particularly hard for the model to learn is sometimes also referred to as group DRO (Sagawa et al., 2020a), as an extension of vanilla distributional robust optimization (DRO) (Ben-Tal et al., 2013; Duchi et al., 2021). Extensions of DRO are mostly discussing the strategies on how to identify the samples considered as minority: Nam et al. (2020) trained two models in parallel, where the "debiased" model focuses on examples not learned by the "biased" model; Lahoti et al. (2020) used an adversary model to identify samples that are challenging to the main model; Liu et al. (2021) proposed to train the model a second time via up-weighting examples that have high training losses during the first time.

When to Use Data-driven or Model-based Approaches?   In many cases both the data and the model can contribute to a model's lack of robustness, hence data-driven and model-based approaches could be combined to further improve a model's robustness. One interesting phenomenon

observed by (Liu et al., 2019) is to attribute models' robustness failures to blind spots in the training data, or the intrinsic learning ability of the model. The authors found that both patterns are possible: in some cases models can be inoculated via being exposed to a small amount of challenging data, similar to the data augmentation approaches mentioned in Section 5.1; on the other hand, some challenging patterns remain difficult which connects to the larger question around generalizability to *unseen* adversarial and counterfactual patterns (Huang et al., 2020; Jha et al., 2020; Joshi and He, 2021), which is relatively under-explored but deserves much attention.

5.3 Inductive-prior-based Approaches

Another thread is to introduce inductive bias (i.e., to regularize the hypothesis space) to force the model to discard some spurious features. This is closely connected to the human-prior-based identification approaches in Section 4.1 as those human-priors can often be used to re-formulate the training objective with additional regularizers. To achieve this goal, one usually needs to first construct a side component to inform the main model about the misaligned features, and then to regularize the main model according to the side component. The construction of this side component usually relies on prior knowledge of what the misaligned features are. Then, methods can be built accordingly to counter the features such as label-associated keywords (He et al., 2019), label-associated text fragments (Mahabadi et al., 2020), and general easy-to-learn patterns of data (Nam et al., 2020). Similarly, Clark et al. (2019, 2020); Utama et al. (2020a,b) propose to *ensemble* with a model explicitly capturing bias, where the main model is trained together with this "bias-only" model such that the main model is discouraged from using biases. More recent work (Xiong et al., 2021) shows the ensemble-based approaches can be further improved via better calibrating the bias-only model. Furthermore, additional regularizers have been introduced for robust fine-tuning over pre-trained models, e.g., mutual-information-based regularizers (Wang et al., 2021a) and smoothness-inducing adversarial regularization (Jiang et al., 2020).

In a broader scope, given that one of the main challenges of domain adaptation is to counter the model's tendency in learning domain-specific spurious features (Ganin et al., 2016), some methods

contributing to domain adaption may have also progressed along the line of our interest, e.g., domain adversarial neural network (Ganin et al., 2016). This line of work also inspires a family of methods forcing the model to learn auxiliary-annotation-invariant representations with a side component (Ghifary et al., 2016; Wang et al., 2017; Rozantsev et al., 2018; Motiian et al., 2017; Li et al., 2018; Wang et al., 2019c; Vernikos et al., 2020).

Despite the diverse concrete ideas introduced, the above is mainly training for small empirical loss across different domains or distributions in addition to forcing the model to be invariant to domain-specific spurious features. As an extension along this direction, invariant risk minimization (IRM) (Arjovsky et al., 2019) introduces the idea of invariant predictors across multiple environments, which was later followed and discussed by a variety of extensions (Choe et al., 2020; Ahmed et al., 2020; Rosenfeld et al., 2021). More recently, Dranker et al. (2021) applied IRM in natural language inference and found that a more naturalistic characterization of the problem setup is needed.

5.4 Causal Intervention

Casual analyses have also been utilized to examine robustness. Srivastava et al. (2020) leverage humans' common sense knowledge of causality to augment training examples with a potential unmeasured variable, and propose a DRO-based approach to encourage the model to be robust to distribution shifts over the unmeasured variables. Balashankar et al. (2021) study the effect of secondary attributes, or confounders, and propose context-aware counterfactuals that take into account the impact of secondary attributes to improve models' robustness. Veitch et al. (2021) propose to learn approximately counterfactual invariant predictors dependent on causal structures of the data, and show it can help mitigate spurious correlations in text classification.

5.5 Connections between Mitigations

Connecting these methods conceptually, we conjecture three different mainstream approaches: one is to leverage the large amount of data by taking advantages of pre-trained models, another is to learn invariant representations or predictors across domains or environments, while most of the rest build upon the prior on what the spurious patterns are and encourage the models to not rely on those patterns. Then the solutions are invented through countering model's learning of these patterns by either data augmentation, reweighting (the minorities), ensemble, inductive-prior design, and causal intervention. Interestingly, statistical work has shown that many of these mitigation methods are optimizing the same robust machine learning generalization error bound (Wang et al., 2021c).

6 Open Questions

In addition to the challenges mentioned above, we list below a few open questions that call for additional research going forward.

Identifying Unknown Robustness Failures Existing identification around robustness failures rely heavily on human priors and error analyses, which usually pre-define a small or limited set of patterns that the model could be vulnerable to. This requires extensive amount of expertise and efforts, and might still suffer from human or subjective biases in the end. How to proactively discover and identify models' unrobust regions automatically and comprehensively remains challenging.

Interpreting and Mitigating Spurious Correlations Interpretability matters for large NLP models, especially key to the robustness and spurious patterns. How can we develop ways to attribute or interpret these vulnerable portions of NLP models and communicate these robustness failures with designers, practitioners, and users? In addition, recent work (Wallace et al., 2019c; Wang et al., 2021d; Zhang et al., 2021) show interpretability methods can be utilized to better understand how a model makes its decision, which in turn can be used to uncover models' bias, diagnose errors, and discover spurious correlations.

Furthermore, the mitigation of spurious correlations often suffers from the trade-off between removing shortcuts and sacrificing model performance (Yang et al., 2020; Zhang et al., 2019a). Additionally, most existing mitigation strategies work in a pipeline fashion where defining and detecting spurious correlations are prerequisites, which might lead to error cascades in this process. How to design end-to-end frameworks for automatic mitigation deserves much attention.

Unified Framework to Evaluate Robustness With a variety of potential spurious patterns in NLP models, it becomes increasingly challenging for developers and practitioners to quickly evaluate the robustness and quality of their models. This calls for more unified benchmarking efforts such as

CheckList (Ribeiro et al., 2020), Reliability Testing (Tan et al., 2021), Robustness Gym (Goel et al., 2021) and Dynabench (Kiela et al., 2021), to facilitate fast and easy evaluation of robustness.

User Centered Measures and Mitigation Instead of passively detecting spurious correlations from a post-processing perspective, how to approach robustness from a user centric perspective needs further investigation. Based on the dual-process models of information processing, humans use two different processing styles (Evans, 2010). One is a quick and automatic style that relies on well-learned information and heuristic cues. The other is a qualitatively different style that is slower, more deliberative, and requires more reflective reasoning. Would these well-learned information and heuristic rules be leveraged to help design better human priors to measure and mitigate spurious correlations? If users or stakeholders are involved in this process, collecting a set of test cases where a system might perform well for the wrong reasons could help design sanity tests.

Connections between Human-like Linguistic Generalization and NLP Generalization Linzen (2020) argue NLP models should behave more like humans to achieve better generalization consistently. It is interesting to note that how humans process information in NLP tasks exactly is still under exploration, and to what extent models should leverage human-knowledge is still a debatable topic.[1] Nonetheless, if we can better understand and utilize the robustness properties in human perception, we can potentially advance models' robustness in a more meaningful way.

7 Conclusion

In this paper, we provided a unifying overview over robustness definitions, evaluations and mitigation strategies in the NLP domain. We also highlighted open challenges in this area to motivate future research, encouraging people to think deeply about more comprehensive benchmarks, transferability and validity of adversarial examples, unified framework to evaluate and improve robustness, user-centered measures and mitigation, and finally how to potentially achieve human-like linguistic generalization more meaningfully.

---

[1]http://www.incompleteideas.net/IncIdeas/BitterLesson.html

References

Alberto Abad, Peter Bell, Andrea Carmantini, and Steve Renals. 2020. Cross lingual transfer learning for zero-resource domain adaptation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6909–6913.

Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron Courville. 2020. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*.

S. Aït-Mokhtar, J.-P. Chanod, and C. Roux. 2002. Robustness beyond shallowness: Incremental deep parsing. *Nat. Lang. Eng.*, 8(3):121–144.

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

Antonios Anastasopoulos, Alison Lui, Toan Q Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *Proceedings of NAACL-HLT*, pages 3070–3080.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Ananth Balashankar, Xuezhi Wang, Ben Packer, Nithum Thain, Ed Chi, and Alex Beutel. 2021. Can we improve model robustness through secondary attribute counterfactuals? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

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

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France. PMLR.

N. Boucher, I. Shumailov, R. Anderson, and N. Papernot. 2022. Bad characters: Imperceptible nlp attacks. In *2022 2022 IEEE Symposium on Security and Privacy (SP)*, pages 773–790, Los Alamitos, CA, USA. IEEE Computer Society.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. 2021a. Hiddencut: Simple data augmentation for natural language understanding with better generalizability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4380–4390.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mix-Text: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021b. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Yangyi Chen, Jin Su, and Wei Wei. 2021c. Multi-granularity textual adversarial attack with behavior cloning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4511–4526, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yo Joong Choe, Jiyeon Ham, and Kyubyong Park. 2020. An empirical study of invariant risk minimization. *arXiv preprint arXiv:2004.05007*.

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 4069–4082, Hong Kong, China. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3031–3045.

Holly Craig and Julie Washington. 2002. Oral language expectations for african american preschoolers and kindergartners. *American Journal of Speech-Language Pathology*, 11:59–70.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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.

Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, and Samson Tan et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation.

Yana Dranker, He He, and Yonatan Belinkov. 2021. Irm - when it works and when it doesn't: A test case of natural language inference. In *Neural Information Processing Systems (NeurIPS)*.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. 2021. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Jonathan St BT Evans. 2010. Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, 21(4):313–326.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *CoRR*, abs/1901.03116.

David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. 2018. From lifestyle vlogs to everyday interactions. In *CVPR*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples (2014). In *International Conference on Learning Representations*.

Alex Graves. 2011. Practical variational inference for neural networks. In *Advances in Neural Informa-*

tion Processing Systems 24, pages 2348–2356. Curran Associates, Inc.

Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *International Conference on Learning Representations*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

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.

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.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020a. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020b. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *CVPR*.

Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. 2021. On robustness of neural semantic parsers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3333–3342, Online. Association for Computational Linguistics.

William Huang, Haokun Liu, and Samuel R. Bowman. 2020. Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 82–87, Online. Association for Computational Linguistics.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136.

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 Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Rohan Jha, Charles Lovering, and Ellie Pavlick. 2020. Does data augmentation improve generalization in nlp?

Robin Jia and Percy Liang. 2017. 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, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. 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 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pretrained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Zhou, and Peter Szolovits. 2020. Is BERT really robust? Natural language attack on text classification and entailment. In *AAAI*.

Nitish Joshi and He He. 2021. An investigation of the (in)effectiveness of counterfactually augmented data.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Subaspect analysis on corpus and system biases in summarization. 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 3324–3335, Hong Kong, China. Association for Computational Linguistics.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Daniel Khashabi, Tushar Khot, and Ashish Sabharwal. 2020. More bang for your buck: Natural perturbation for robust question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 163–170, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

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

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, Online. Association for Computational Linguistics.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17,

page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Kong Joo Lee, Cheol Jung Kweon, Jungyun Seo, and Gil Chang Kim. 1995. A robust parser based on syntactic information. In *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.

Qi Lei, Lingfei Wu, Pin-Yu Chen, Alex Dimakis, Inderjit S. Dhillon, and Michael J Witbrock. 2019. Discrete adversarial attacks and submodular optimization with applications to text classification. In *Proceedings of Machine Learning and Systems*, volume 1, pages 146–165.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. Textbugger: Generating adversarial text against real-world applications. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Pro-

ceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5210–5217. Association for Computational Linguistics.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. 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 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8706–8716. Association for Computational Linguistics.

Gabriel Marzinotto, Géraldine Damnati, Frédéric Béchet, and Benoît Favre. 2019. Robust semantic parsing with adversarial learning for domain generalization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 166–173, Minneapolis, Minnesota. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. 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.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. 2016. Launch and iterate: Reducing prediction churn. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3179–3187. Curran Associates, Inc.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *Proceedings of the International Conference on Learning Representations*.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020a. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020b. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 3.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, Atlanta, Georgia, USA. PMLR.

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.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust language modeling. In *EMNLP/IJCNLP (1)*.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. 2015. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Marcelo Prates, Pedro Avelar, and Luís Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. 2021. The risks of invariant risk minimization. In *International Conference on Learning Representations*.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2021. Tailor: Generating and perturbing text with semantic controls.

Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. 2018. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020a. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020b. An investigation of why overparameterization exacerbates spurious correlations.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2021. Semantics altering modifications for evaluating comprehension in machine reading. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 13762–13770. AAAI Press.

Leo Schwinn, René Raab, An Nguyen, Dario Zanca, and Bjoern Eskofier. 2021. Exploring misclassifications of robust neural networks to enhance adversarial attacks.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2020. Robustness to modification with shared words in paraphrase identification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 164–171, Online. Association for Computational Linguistics.

Walter Simoncini and Gerasimos Spanakis. 2021. SeqAttack: On adversarial attacks for named entity recognition. In *Proceedings of the 2021 Conference*

4583

on Empirical Methods in Natural Language Processing: System Demonstrations, pages 308–318, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733, Online. Association for Computational Linguistics.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9109–9119. PMLR.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Samson Tan and Shafiq Joty. 2021. Code-mixing on sesame street: Dawn of the adversarial polyglots. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3596–3616, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020a. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020b. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2019.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*.

Giorgos Vernikos, Katerina Margatina, Alexandra Chronopoulou, and Ion Androutsopoulos. 2020. Domain adversarial fine-tuning as an effective regularizer. *CoRR*, abs/2009.13366.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Ya-mada, and Jordan Boyd-Graber. 2019b. Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019c. AllenNLP interpret: A framework for explaining predictions of NLP 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): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150, Online. Association for Computational Linguistics.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021a. Infobert: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021b. Adversarial glue: A multi-task benchmark for robustness evaluation of language models.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. 2019a. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518.

Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. 2019b. Learning robust representations by projecting superficial statistics out. In *7th International Conference on Learning Representations, ICLR 2019*.

Haohan Wang, Zeyi Huang, Hanlin Zhang, and Eric Xing. 2021c. Toward learning human-aligned cross-domain robust models by countering misaligned features.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo, ICME 2017*, pages 949–954. IEEE Computer Society.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. 2020a. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694.

Haohan Wang, Zhenglin Wu, and Eric P. Xing. 2019c. Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications. In *Biocomputing 2019: Proceedings of the Pacific Symposium*, pages 54–65.

Haohan Wang, Peiyan Zhang, and Eric P Xing. 2020b. Word shape matters: Robust machine translation with visual embedding. *arXiv preprint arXiv:2010.09997*.

Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019d. Adversarial domain adaptation for machine reading comprehension. 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)*.

Tianlu Wang, Xuezhi Wang, Yao Qin, Ben Packer, Kang Li, Jilin Chen, Alex Beutel, and Ed Chi. 2020c. CAT-gen: Improving robustness in NLP models via controlled adversarial text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5141–5146, Online. Association for Computational Linguistics.

Tianlu Wang, Diyi Yang, and Xuezhi Wang. 2021d. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*.

Xiaosen Wang, Hao Jin, and Kun He. 2019e. Natural language adversarial attacks and defenses in word level. *arXiv preprint arXiv:1909.06723*.

Zhao Wang and Aron Culotta. 2020a. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Zhao Wang and Aron Culotta. 2020b. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *AAAI*.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020d. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online. Association for Computational Linguistics.

Garrett Wilson and Diane J. Cook. 2020. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5).

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving

models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Ruibin Xiong, Yimeng Chen, Liang Pang, Xueqi Cheng, Zhi-Ming Ma, and Yanyan Lan. 2021. Uncertainty calibration for ensemble-based debiasing methods. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. 2020. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems*, volume 33, pages 8588–8601. Curran Associates, Inc.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Liping Yuan, Xiaoqing Zheng, Yi Zhou, Cho-Jui Hsieh, and Kai-Wei Chang. 2021. On the transferability of adversarial attacks against neural text classifier. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1612–1625, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

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.

Dejiao Zhang, Ramesh Nallapati, Henghui Zhu, Feng Nan, Cicero dos Santos, Kathleen McKeown, and Bing Xiang. 2020a. Unsupervised domain adaptation for cross-lingual text labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3527–3536.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019a. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020c. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. 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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018b. Generating natural adversarial examples. In *ICLR*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for language understanding. In *ICLR*.