Context-aware Adversarial Training for Name Regularity Bias in Named Entity Recognition

Abbas Ghaddar, Philippe Langlais[†], Ahmad Rashid, and Mehdi Rezagholizadeh

Huawei Noah's Ark Lab, Montreal Research Center, Canada
[†]RALI/DIRO, Université de Montréal, Canada
`abbas.ghaddar@huawei.com, felipe@iro.umontreal.ca`
`ahmad.rashid@huawei.com, mehdi.rezagholizadeh@huawei.com`

Abstract

In this work, we examine the ability of NER models to use contextual information when predicting the type of an ambiguous entity. We introduce NRB, a new testbed carefully designed to diagnose **N**ame **R**egularity **B**ias of NER models. Our results indicate that all state-of-the-art models we tested show such a bias; BERT fine-tuned models significantly outperforming feature-based (LSTM-CRF) ones on NRB, despite having comparable (sometimes lower) performance on standard benchmarks.

To mitigate this bias, we propose a novel model-agnostic training method that adds *learnable* adversarial noise to some entity mentions, thus enforcing models to focus more strongly on the contextual signal, leading to significant gains on NRB. Combining it with two other training strategies, data augmentation and parameter freezing, leads to further gains.

1 Introduction

Recent advances in language model pre-training (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019) have greatly improved the performance of many Natural Language Understanding (NLU) tasks. Yet, several studies (McCoy et al., 2019; Clark et al., 2019; Utama et al., 2020b) revealed that state-of-the-art NLU models often make use of surface patterns in the data that do not generalize well. Named-Entity Recognition (NER), a downstream task that consists in identifying textual mentions and classifying them into a predefined set of types, is no exception.

The robustness of modern NER models has received considerable attention recently (Mayhew et al., 2019; Mayhew et al., 2020; Agarwal et al., 2020a; Zeng et al., 2020; Bernier-Colborne and Langlais, 2020). Name Regularity Bias (Lin et al.,

2020; Agarwal et al., 2020b; Zeng et al., 2020) in NER occurs when a model relies on a signal coming from the entity name, and disregards evidence within the local context. Figure 1 shows examples where state-of-the-art models (Peters et al., 2018; Akbik et al., 2018; Devlin et al., 2019) fail to exploit contextual information. For instance, the entity *Gonzales* in the first sentence of the figure is wrongly recognized as a *person*, while the context clearly signals that it is a *location* (city).

To better highlight this issue, we propose NRB, a testbed designed to accurately diagnose name regularity bias of NER models by harvesting natural sentences from Wikipedia that contain challenging entities, such as those in Figure 1. This is different from previous work that evaluated models on artificial data obtained by either randomizing (Lin et al., 2020) or substituting entities by ones from a pre-defined list (Agarwal et al., 2020a). NRB is compatible with any annotation scheme, and is intended to be used as an auxiliary validation set.

We conduct experiments with the *feature-based* LSTM-CRF architecture (Peters et al., 2018; Akbik et al., 2018) and the BERT (Devlin et al., 2019) *fine-tuning* approach trained on standard benchmarks. The best LSTM-based model we tested is able to correctly predict 38% of the entities in NRB. BERT-based models are performing much better (+37%), even if they (slightly) underperform on in-domain development and test sets. This mismatch in performance between NRB and standard benchmarks indicates that context awareness of models is not rewarded by existing benchmarks, thus justifying NRB as an additional validation set.

We further propose a novel architecture-agnostic adversarial training procedure (Miyato
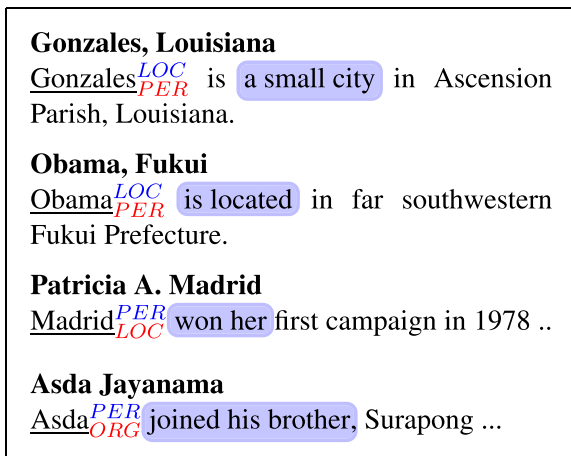
586

Gonzales, Louisiana

Gonzales$_{PER}^{LOC}$ is a small city in Ascension Parish, Louisiana.

Obama, Fukui

Obama$_{PER}^{LOC}$ is located in far southwestern Fukui Prefecture.

Patricia A. Madrid

Madrid$_{LOC}^{PER}$ won her first campaign in 1978 ..

Asda Jayanama

Asda$_{ORG}^{PER}$ joined his brother, Surapong ...

Figure 1: Examples extracted from Wikipedia (title in bold) that illustrate name regularity bias in NER. Entities of interest are underlined, gold types are in blue superscript, model predictions are in red subscript, and context information is highlighted in purple. Models used in this study disregard contextual information and rely instead on some signal from the named-entity itself.

et al., 2016) in which *learnable* noise vectors are added to named-entity words, weakening their signal, thus encouraging the model to pay more attention to contextual information. Applying it to both feature-based LSTM-CRF and fine-tuned BERT models leads to consistent gains on NRB (+13 points) while maintaining the same level of performance on standard benchmarks.

The remainder of the paper is organized as follows. We discuss related works in Section 2. We describe how we built NRB in Section 3, and its use in diagnosing named-entity bias of state-of-the-art models in Section 4. In Section 5, we present a novel adversarial training method that we compare and combine with two simpler ones. We further analyze these training methods in Section 6, and conclude in Section 7.

2 Related Work

Robustness and out-of-distribution generalization has always been a persistent concern in deep learning applications such as computer vision (Szegedy et al., 2013; Recht et al., 2019), speech processing (Seltzer et al., 2013; Borgholt et al., 2020), and NLU (Søgaard, 2013; Hendrycks and Gimpel, 2017; Ghaddar and Langlais, 2017; Yaghoobzadeh et al., 2019; Hendrycks et al., 2020). One key challenge behind this issue in

NLU is the tendency of models to quickly leverage surface form features and annotation artifacts (Gururangan et al., 2018), which is often referred to as dataset biases (Dasgupta et al., 2018; Shah et al., 2020). We discuss related works along two axes: diagnosis and mitigation.

2.1 Diagnosing Biais

A growing number of studies (Zellers et al., 2018; Poliak et al., 2018; Geva et al., 2019; Utama et al., 2020b; Sanh et al., 2020) are showing that NLU models rely heavily on spurious correlations between output labels and surface features (e.g., keywords, lexical overlap), impacting their generalization performance. Therefore, considerable attention has been paid to design diagnostic benchmarks where models relying on bias would perform poorly. For instance, HANS (McCoy et al., 2019), FEVER Symmetric (Schuster et al., 2019), and PAWS (Zhang et al., 2019) are benchmarks that contain counterexamples to well-known biases in the training data of textual entailment (Williams et al., 2017), fact verification (Thorne et al., 2018), and paraphrase identification (Wang et al., 2018), respectively.

Naturally, many entity names have a strong correlation with a single type (e.g., <Gonzales, PER> or <Madrid, LOC>). Recent works have noted that over-relying on entity name information negatively impacts NLU tasks. Balasubramanian et al. (2020) found that substituting named-entities in standard test sets of natural language inference, coreference resolution, and grammar error correction has a negative impact on those tasks. In political claims detection (Padó et al., 2019), Dayanik and Padó (2020) show that claims made by frequently occurring politicians in the training data are better recognized than those made by less frequent ones.

Recently, Zeng et al. (2020) and Agarwal et al. (2020b) conducted two separate analyses on the decision making mechanism of NER models. Both works found that context tokens do contribute to system performance, but that entity names play a major role in driving high performances. Agarwal et al. (2020a) reported a performance drop in NER models when entities in standard test sets are substituted with other ones pulled from pre-defined lists. Concurrently, Lin et al. (2020) conducted an empirical analysis on the robustness of NER models in the open domain

587

scenario. They show that models are biased by strong entity name regularity, and train\test overlap in standard benchmarks. They observe a drop in performance of 34% when entity mentions are randomly replaced by other mentions.

The aforementioned studies certainly demonstrate name regularity bias. Still, in many cases the entity mention is the only key to infer its type, as in ''*James won the league*''. Thus, randomly swapping entity names, as proposed by Lin et al. (2020), typically introduces false positive examples, which obscures observations. Furthermore, creating artificial word sequences introduces a mismatch between the pre-training and the fine-tuning phases of large-scale language models.

NER is also challenging because of compounding factors such as entity boundary detection (Zheng et al., 2019), rare words and emerging entities (Strauss et al., 2016), document-level context (Durrett and Klein, 2014), capitalization mismatch (Mayhew et al., 2019), unbalance datasets (Nguyen et al., 2020), and domain shift (Alvarado et al., 2015; Augenstein et al., 2017). It is unclear to us how randomizing mentions in a corpus, as proposed by Lin et al. (2020), is interfering with these factors.

NRB gathers genuine entities that appear in natural sentences extracted from Wikipedia. Examples are selected so that entity boundaries are easy to identify, and their types can be inferred from the local context, thus avoiding compounding many factors responsible for lack of robustness.

2.2 Mitigating Bias

The prevailing approach to address dataset bias consists in adjusting the training loss for biased examples. A number of recent studies (Clark et al., 2019; Belinkov et al., 2019; He et al., 2019; Mahabadi et al., 2020; Utama et al., 2020a) proposed to train a shallow model that exploits manually designed biased features. A main model is then trained in an ensemble with this pre-trained model, in order to discourage the main model from adopting the naive strategy of the shallow one.

Adversarial training (Miyato et al., 2016) is a regularization method that has been shown to improve not only robustness (Ebrahimi et al., 2018; Bekoulis et al., 2018), but also generalization (Cheng et al., 2019; Zhu et al., 2019) in NLU. It builds on the idea of adding adversarial examples (Goodfellow et al., 2014; Fawzi et al.,

2016) to the training set, that is, small perturbations of the data that can change the prediction of a classifier. These perturbations for NLP tasks are done at the token embedding level and are norm bounded. Typically, adversarial training algorithms can be defined as a *minmax* optimization problem wherein the adversarial examples are generated to maximize the loss, while the model is trained to minimize it.

Belinkov et al. (2019) used adversarial training to mitigate the hypothesis-only bias in textual entailment models. Clark et al. (2020) adversarially trained a low and a high capacity model in an ensemble in order to ensure that the latter model is focusing on patterns that should generalize better. Dayanik and Padó (2020) used an extra adversarial loss in order to encourage a political claims detection model to learn more from samples with infrequent politician names. Le Bras et al. (2020) proposed an adversarial technique to filter-out biased examples from training material. Models trained on the filtered datasets show improved out-of-distribution performances on various computer vision and NLU tasks.

Data augmentation is another strategy for enhancing robustness. It was successfully used in Min et al. (2020) and Moosavi et al. (2020) to improve textual entailment performances on the HANS benchmark. The former approach proposes to append original training sentences with their corresponding predicate-arguments triplets generated by a semantic role labelling tagger; while the latter generates new examples by applying syntactic transformations to the original training instances.

Zeng et al. (2020) created new examples by randomly replacing an entity by another one of the same type that occurs in the training data. New examples are considered valid if the type of the replaced entity is correctly predicted by a NER model trained on the original dataset. Similarly, Dai and Adel (2020) explored different entity substitution techniques for data augmentation tailored to NER. Both studies conclude that data augmentation techniques based on entity substitution improves the overall performances on low resource biomedical NER.

Studies discussed above have the potential to mitigate name regularity bias of NER models. Still, we are not aware of any dedicated work that shows it is so. In this work, we propose ways of

mitigating name regularity bias for NER, including an elaborate adversarial method that forces the model to capture more signal from the context. Our methods do not require an extra training stage, or to manually characterize biased features. They are therefore conceptually simpler, and can potentially be combined to any of the discussed techniques. Furthermore, our proposed methods are effective under both low and high resource settings.

3 The NRB Benchmark

NRB is a diagnosing testbed exclusively dedicated to name regularity bias in NER. To this end, it gathers named-entities that satisfy 4 criteria:

1. Must be real-world entities within natural sentences → We select sentences from Wikipedia articles.

2. Must be compatible with any annotation scheme → We restrict our focus on the 3 most common types found in NER benchmarks: *person*, *location*, and *organization*.

3. Boundary detection (segmentation) should not be a bottleneck → We only select single word entities that start with a capital letter.

4. Supporting evidences of the type must be restricted to local context only (a window of 2 to 4 tokens) → We developed a primitive *context-only* tagger to filter-out entities with no close-context signal.

The strategy used to gather examples in NRB is illustrated in Figure 2. We first select Wikipedia articles that are listed in a disambiguation page. Disambiguation pages group different topics that could be referred to by the same query term.[1] The query term *Bromwich* in Figure 2 has its own disambiguation page that contains a link to the city of *West Bromwich*, *West Bromwich Albion Football Club*, and *Kenny Bromwich* the rugby league player.

We associate each article in a disambiguation page to the entity type found in its corresponding Freebase page (Bollacker et al., 2008), considering only articles whose Freebase type can be mapped to a person, a location, or an organization. We assume that occurrences of the query term

---

Disambiguation page	
Bromwich (disambiguation)	
Query term	Bromwich
Wikipedia article	*Kenny Bromwich*
Freebase type	*PER*
Sentence	
Round 5 of the 2013 NRL season <u>Bromwich</u> made his NRL debut for the Melbourne Storm	
Taggers	
weak supervision	ORG (**confidence**: 0.97)
context-only	PER: 0.58, ORG: 0.30, LOC: 0.12

Figure 2: Selection of a sentence in NRB.

within the article are of this type. This assumption was found accurate in previous work on Wikipedia distant supervision for NER (Ghaddar and Langlais, 2016, 2018). The sentence in our example is extracted from the *Kenny Bromwich* article, whose Freebase type can be mapped to a person. Therefore, we assume *Bromwich* in this sentence to be a person.

To decide whether a sentence containing a query term is worth being included in NRB, we rely on two NER taggers. One is a popular NER system that provides a confidence score to each prediction, and that acts as a *weak superviser*, the other is a *context-only* tagger we designed specifically (see Section 3.1) to detect entities with a strong signal from their local context. A sentence is selected if the query term is incorrectly labeled with high confidence (score > 0.85) by the former tagger, while the latter one labels it correctly with high confidence (a gap of at least 0.25 in probability between the first and second predicted types). This is the case of the sentence in Figure 2, where *Bromwich* is incorrectly labeled as an organization by the weak supervision tagger, however correctly labeled as a person by the context-only tagger.

3.1 Implementation

We used the Stanford CoreNLP (Manning et al., 2014) tagger as our weak supervision tagger and developed a simple yet efficient method to build a context-only tagger. For this, we first applied the Stanford tagger to the entire Wikipedia dump and replaced all entity mentions identified by their tag. Then, we train a 5-gram language model on the resulting corpus using kenLM (Heafield, 2011).

---

[1] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Disambiguation_pages.

> *Obama* is located in far southwestern Fukui Prefecture.
> <?> is located in far southwestern Fukui Prefecture.
> {*LOC*: 0.61, *ORG*: 0.28, *PER*: 0.11}

Figure 3: Illustration of a language model used as a context-only tagger.

Figure 3 illustrates how this model is deployed as an entity tagger: The mention is replaced by an empty slot and the language model is queried for each type. We rank the tags using the perplexity score given by the model to the resulting sentences, then we normalize those scores to get a probability distribution over types.

We downloaded the Wikipedia dump of June 2020, which contains 30k disambiguation pages. These pages contain links to 263k articles, where only 107k (40%) of them have a type in Freebase that can be mapped to the 3 types of interest. The Stanford tagger identified 440k entities that match the query term of the disambiguation pages. The thresholds discussed previously were chosen to select around 5000 of the most challenging examples in terms of name regularity bias. This figure aligns with the number of entities present in the test set of the well-studied CoNLL benchmark (Tjong Kim Sang and De Meulder, 2003).

We assessed the annotation quality by asking a human to filter out noisy examples. A sentence was removed if it contains an annotation error, or if the type of the query term cannot be inferred from the local context. Only 1.3% of the examples where removed, which confirms the accuracy of our automatic procedure. NRB is composed of 5275 examples, and each sentence contains a single annotation (see Figure 1 for examples).

3.2 Control Set (WTS)

In addition to NRB, we collected a set of domain control sentences—called WTS for WITNESS—that contain the very same query terms selected in NRB, but that were correctly labeled by both the Stanford (score > 0.85) and the context-only taggers. We selected examples with a small gap (< 0.1) between the first and second ranked type assigned to the query term by the latter tagger. Thus, examples in WTS should be easy to tag. For example, because *Obama* the Japanese city

(see Figure 3) is selected among the query terms in NRB, we added an instance of *Obama* the president.

Performing poorly on such examples[2] indicates a domain shift between NRB (Wikipedia) and whatever dataset a model is trained on (we call it the in-domain corpus). WTS is composed of 5192 sentences that have also been manually checked.

4 Diagnosing Bias

4.1 Data

To be comparable with state-of-the-art models, we consider two standard benchmarks for NER: CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) and ONTONOTES 5.0 (Pradhan et al., 2012), which include 4 and 18 types of named-entities, respectively. ONTONOTES is 4 times larger than CoNLL, and both benchmarks mainly cover the news domain. We run experiments on the official train/dev/test splits, and report mention-level F1 scores, following previous work. Since in NRB, there is only one entity per sentence to annotate, a system is evaluated on its ability to correctly identify the boundaries of this entity and its type. When we train on ONTONOTES (18 types) and evaluate on NRB (3 types), we perform type mapping using the scheme of Augenstein et al. (2017).

4.2 Systems

Following (Devlin et al., 2019), we term all approaches that learn the encoder from scratch as *feature-based*, as opposed to the ones that fine-tune a pre-trained model for the downstream task. We conduct experiments using 3 *feature-based* and 2 *fine-tuning* approaches for NER:

- **Flair-LSTM** An LSTM-CRF model that uses FLAIR (Akbik et al., 2018) contextualized embeddings as main features.

- **ELMo-LSTM** The LSTM-CRF tagging model of Peters et al. (2018) that uses ELMo contextualized embeddings at the input layer.

- **BERT-LSTM** Similar to the previous model, but replacing ELMo by a representation gathered from the last four layers of BERT.

- **BERT-base** The fine-tuning approach proposed by Devlin et al. (2019) using the BERT-base model.

---

[2]That is, a system that fail to tag *Obama* the president as a person.

590

Model	CoNLL				OntoNotes			
	Dev	Test	NRB	WTS	Dev	Test	NRB	WTS
Feature-based								
Flair-LSTM	-	**93.03**	27.56	**99.58**	-	89.06	33.67	93.98
ELMo-LSTM	96.69	92.47	31.65	98.24	88.31	89.38	34.34	94.90
BERT-LSTM	95.94	91.94	38.34	98.08	86.12	87.28	43.07	92.04
Fine-tuning								
BERT-base	96.18	92.19	75.54	98.67	87.23	88.19	75.34	94.22
BERT-large	**96.90**	92.86	**75.55**	98.51	**89.26**	**89.93**	**75.41**	**95.06**

Table 1: Mention level F1 scores of models on CoNLL and OntoNotes, as well as on NRB and WTS.

- **BERT-large** The fine-tuning approach using the BERT-large model.

We used Flair-LSTM off-the-shelf,[3] and re-implemented other approaches using the default settings proposed in the respective papers. For our reimplementations, we used early stopping based on performance on the development set, and report average performance over 5 runs. For BERT-based solutions, we adopt spanBERT (Joshi et al., 2020) as a backbone model because it was found by Li et al. (2020) to perform better on NER.

4.3 Results

Table 1 shows the mention level F1 score of the systems considered. Flair-LSTM and BERT-large are the best performing models on in-domain test sets, the maximum gap with other models being 1.1 and 2.7 on CoNLL and OntoNotes respectively. These figures are in line with previous work. What is more interesting is the performance on NRB. Feature-based models do poorly, Flair-LSTM underperforms compared to other models (F1 score of 27.6 and 33.7 when trained on CoNLL and OntoNotes respectively). Fine-tuned BERT models clearly perform better (around 75), but far from in-domain results (92.9 and 89.9 on CoNLL and OntoNotes, respectively). Domain shift is not a reason for those results, since the performances on WTS are rather high (92 or higher). Furthermore, we found that the boundary detection (segmentation) performance on NRB is above 99.2% across all settings. Because errors made on NRB are neither due to segmentation nor to domain shift, they must be imputed to name regularity bias of models.

It is worth noting that BERT-LSTM outperforms ELMo-LSTM on NRB, despite underperforming on in-domain test sets. This may be

because BERT was pre-trained on Wikipedia (same domain of NRB), while ELMo embeddings were trained on the One Billion Word corpus (Chelba et al., 2014). Also, we observe that switching from BERT-base to BERT-large, or training on 4 times more data (CoNLL versus OntoNotes) does not help on NRB. This suggests that name regularity bias is neither a data nor a model capacity issue.

4.4 Feature-based vs. Fine-tuning

In this section, we analyze reasons for the drastic superiority of fined-tuned models on NRB. First, the large gap between BERT-LSTM and BERT-base on NRB suggests that this is not related to the representations being used at the input layer.

Second, we tested several configurations of ELMo-LSTM where we scale up the number of LSTM layers and hidden units. We observed a degradation of performance on dev, test, and NRB sets, mostly due to over-parameterized models. We also trained 9-, 6-, and 4-layer BERT-base models,[4] and still noticed a large advantage of BERT models on NRB.[5] This suggests that the higher capacity of BERT alone cannot explain all the gains.

Third, since by design, evidence on the entity type in NRB resides within the local context, it is unlikely that gains on this set come from the ability of Transformers (Vaswani et al., 2017) to better handle long dependencies than LSTM (Hochreiter and Schmidhuber, 1997). To further validate this statement, we fine-tuned BERT models with randomly initialized weights, except the embedding layer. We noticed that this time, the performances

---

[3]https://github.com/flairNLP/flair.

[4]We used early exit (Xin et al., 2020) at the kth layer.

[5]The 4-layer model has 53M parameters and performs 52% on NRB.

on NRB fall into the same range of those of feature-based models, and a drastic decrease (12%–15%) on standard benchmarks. These observations are in keeping with results from Hendrycks et al. (2020) on the out-of-distribution robustness of fine-tuning pre-trained transformers, and also confirms observations made by Agarwal et al. (2020b).

From these analyses, we conclude that the Masked Language Model (MLM) objective (Devlin et al., 2019) that the BERT models were pre-trained with is a key factor driving superior performance of the *fine-tuned* models on NRB. In most cases, the target word is masked or randomly selected, therefore the model must rely on the context to predict the correct target, which is what a model should do to correctly predict the type of entities in NRB. We think that in fine-tuning, training for a few epochs with a small learning rate helps the model to preserve the *contextual behavior* induced by the MLM objective.

Nevertheless, fine-tuned models recording at best an F1 score of 75.6 on NRB do show some name regularity bias, and fail to capture useful local contextual information.

5 Mitigating Bias

In this section, we investigate training procedures that are designed to enhance the contextual awareness of a model, leading to better performance on NRB without impacting in-domain performance. These training procedures are not supposed to use any external data. In fact, NRB is only used as a diagnosing corpus, once the model is trained. We propose 3 training procedures that can be combined, two of them are architecture-agnostic, and one is specific to *fine-tuning* BERT.

5.1 Entity Masking

Inspired by the masking strategy applied during the pre-training phase of BERT, we propose a data augmentation approach that introduces a special [MASK] token in some of the training examples. Specifically, we search for entities in the training material that are preceded or followed by 3 non-entity words. This criterion applies to 35% and 39% of entities in the training data of CoNLL and OntoNotes, respectively. For each such entity, we create a new training example (new sentence) by replacing the entity by [MASK], thus forcing

the model to infer the type of masked tokens from the context. We call this procedure *mask*.

5.2 Parameter Freezing

Another simple strategy, specific to *fine-tuning* BERT, consists of freezing part of the network. More precisely, we freeze the bottom half of BERT, including the embedding layer. The intuition is to preserve part of the *predicting-by-context* mechanism that BERT has acquired during the pre-training phase. This training procedure is expected to enforce the contextual ability of the model, thus adding to our analysis on the critical role of the MLM objective in pre-training BERT. We name this method *freeze*.

5.3 Adversarial Noise

We propose an adversarial learning algorithm that makes entity type patterns in the input representation less reliable for the model, thus enforcing it to rely more aggressively on the context. To do so, we add a learnable adversarial noise vector (only) to the input representation of entities. We refer to this method as *adv*.

Let $T = \{t_1, t_2, \ldots, t_K\}$ be a predefined set of types such as PER, LOC, and ORG in our case. Let $x = x_1, x_2, \ldots, x_n$ be the input sequence of length n, $y = y_1, y_2, \ldots, y_n$ be the gold label sequence following the IOB[6] tagging scheme, and $y' = y'_1, y'_2, \ldots, y'_n$ be a sequence obtained by adding noise to y at the mention-level, that is, by randomly replacing the type of mentions in y with some noisy type sampled from T.

Let $\mathcal{Y}_{ij}(t) = y_i, \ldots, y_j$ be a mention of type $t \in T$, spanning the sequence of indices i to j in y. We derive a noisy mention $\mathcal{Y}'_{ij}$ in y' from $\mathcal{Y}_{ij}(t)$ as follows:

$$\mathcal{Y}'_{ij} = \begin{cases} \mathcal{Y}_{ij}(t') & p \sim U(0,1) \le \lambda \\ & t' \sim \underset{\gamma \in T \backslash \{t\}}{\mathrm{Cat}}(\gamma | \xi = \frac{1}{K-1}) \\ \mathcal{Y}_{ij}(t) & \text{otherwise} \end{cases}$$

where λ is a threshold parameter, $U(0,1)$ refers to the uniform distribution in the range [0,1], $\mathrm{Cat}(\gamma | \xi = \frac{1}{K-1})$ is the categorical distribution whose outcomes are equally likely with the probability of ξ, and the set $T \backslash \{t\} = \{t' : t' \in T \wedge t' \ne t\}$ stands for the set T excluding type t.

---

[6]Naturally applies to other schemes, such as BILOU that Ratinov and Roth (2009) found more informative.
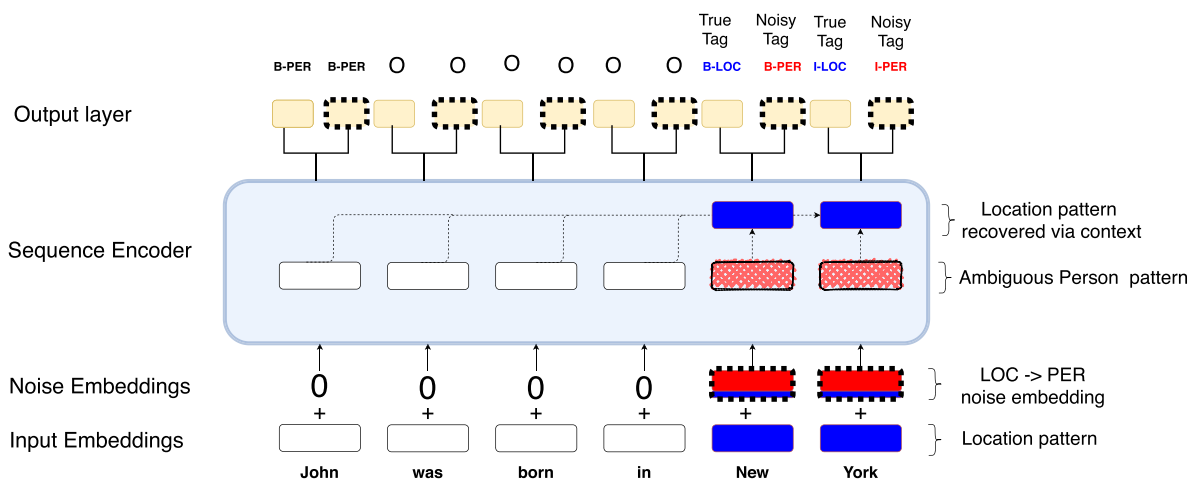
Figure 4: Illustration of our adversarial method applied on the entity *New York*. First, we generate a noisy type (PER), and then add a learnable noise embedding (LOC→PER) to the input representation of that entity. This will make entity patterns (hashed rectangles) unreliable for the model, hence forcing it to collect evidences (dotted arrow) from the context. The noise embedding matrix and the noise label projection layer weights (dotted rectangle) are trained independently from the model parameters.

The above procedure only applies to the entities that are preceded or followed by 3 context words. For instance, in Figure 4, we produce a noisy type for *New York* (PER), but not for *John* ($p > \lambda$). Also, note that we generate a different sequence y' from y at each training epoch.

Next, we define a learnable noisy embedding matrix $E' \in \mathbb{R}^{m \times d}$ where $m = |T| \times (|T| - 1)$ is the number of valid type switching possibilities, and d is the dimension of the input representations of x. For each token with a noisy label, we add the corresponding noisy embedding to its input representation. For other tokens, we simply add a zero vector of size d. As depicted in Figure 4, the noisy type of the entity *New York* is PER, therefore we add the noise embedding at index $LOC \rightarrow PER$ to its input representation.

Then, the input representation of the sequence is fed to an encoder followed by an output layer, such as LSTM-CRF in Peters et al. (2018), or BERT-Softmax in Devlin et al. (2019). First, we extend the aforementioned models by generating an extra logit f' using a projection layer parametrized by W' and followed by a softmax function. As shown in Figure 4, for each token the model produces two logits relative to the true and noisy tags. Then, we train the entire model to minimize two losses: $L_{true}(\theta)$ and $L_{noisy}(\theta')$, where θ is the original set of parameters and $\theta' = \{E', W'\}$ is the extra set we added (dotted boxes in Figure 4). $L_{true}(\theta)$ is

the regular loss on the true tags, while $L_{noisy}(\theta')$ is the loss on the noisy tags defined as follows:

$$L_{\text{noisy}}(\theta') = \sum_{i=1}^{n} \mathbb{1}\left(y'_i \neq y_i\right) \text{CE}\left(f'_i, y'_i\right)$$

where CE is the cross-entropy loss function. Both losses are minimized using gradient descent. It is worth mentioning that λ is the only hyper-parameter of our *adv* method. It controls how often noisy embeddings are added during training. Higher values of λ increase the amount of uncertainty around salient patterns in the input representation of entities, hence preventing the model from overfitting those patterns, and therefore pushing it to rely more on context information. We tried values of λ between 0.3 and 0.9, and found $\lambda = 0.8$ to be the best one based on CoNLL and ONTONOTES development sets.

5.4 Results

We trained models on CoNLL and ONTONOTES, and evaluated them on their respective TEST set.[7] Recall that NRB and WTS are only used as auxiliary diagnosing sets. Table 2 shows the impact of our training methods when fine-tuning the BERT-large model (the one that performs best on NRB).

First, we observe that each training method significantly improves the performance on NRB.

---

[7]Performances on DEV show very similar trends.

Method	CoNLL			OntoNotes		
	Test	NRB	WTS	Test	NRB	WTS
BERT-lrg	92.8	75.6	**98.6**	89.9	75.4	95.1
+mask	**92.9**	82.9	98.4	89.8	77.3	**96.5**
+freeze	92.7	83.1	98.4	89.9	79.8	96.0
+adv	92.7	86.1	98.3	**90.1**	85.8	95.2
+f&m	92.8	85.5	97.8	89.9	80.6	95.9
+a&m	92.8	87.7	98.1	89.7	87.6	95.9
+a&f	92.7	88.4	98.2	90.0	88.1	95.7
+a&m&f	92.8	**89.7**	97.9	89.9	**88.8**	95.6

Table 2: Impact of training methods on BERT-large models fine-tuned on CoNLL or OntoNotes.

Method	CoNLL			OntoNotes		
	Test	NRB	WTS	Test	NRB	WTS
E-LSTM	**92.5**	31.7	**98.2**	**89.4**	34.3	94.9
+mask	92.4	40.8	97.5	89.3	38.8	**95.3**
+adv	92.4	42.4	97.8	89.4	40.7	95.0
+a&m	92.4	**45.7**	96.8	89.3	**46.6**	93.7

Table 3: Impact of training methods on the ELMo-LSTM trained on CoNLL or OntoNotes.

Adding adversarial noise is notably the best performing method on NRB, with an additional gain of 10.5 and 10.4 F1 points over the respective baselines. On the other hand, we observe minor variations on in-domain test sets, as well as on WTS. The paired sample t-test (Cohen, 1996) confirms that these variations are not statistically significant ($p > 0.05$). After all, the number of decisions that differ between the baseline and the best model on a given in-domain set is less than 20.

Second, we observe that combining methods always leads to improvements on NRB; the best configuration being when we combine all 3 methods. It is interesting to note that combining training methods leads to a performance on NRB which does not depend much on the training set used: CoNLL (89.7) and OntoNotes (88.8). This suggests that name regularity bias is a modeling issue, and not the effect of factors such as training data size, domain, or type granularity.

In order to validate that our training methods are not specific to the *fine-tuning* approach, we replicated the same experiments with the ELMo-LSTM. Table 3 shows the performance of the *mask* and *adv* procedures (the *freeze* method does not apply here). The results are in line with those observed with BERT-large: significant

gains on NRB of 14 and 12 points for CoNLL and OntoNotes models, respectively, and no statistically significant changes on in-domain test sets. Again, combining training methods leads to systematic gains on NRB (13 points on average). Differently from fine-tuning BERT, we observe a slight drop in performance of 1.2% on WTS when both methods are used.

The performance of ELMo-LSTM on NRB does not rival the one obtained by fine-tuning the BERT-large model, which confirms that BERT is a key factor to enhance robustness, even if in-domain performance is not necessarily rewarded (McCoy et al., 2019; Hendrycks et al., 2020).

6 Analysis

So far, we have shown that state-of-the-art models do suffer from name regularity bias, and we proposed model-agnostic training methods that are able to mitigate this bias to some extent. In Section 6.1, we provide further evidence that our training methods force the BERT-large model to better concentrate on contextual cues. In Section 6.2, we replicate the evaluation protocol of Lin et al. (2020) in order to clear out the possibility that our training methods are only valid on NRB. Last, we perform extensive experiments on name regularity bias under low resource (Section 6.3) and multilingual (Section 6.4) settings.

6.1 Attention Heads

We leverage the attention map of BERT to better understand how our method enhances context encoding. To this end, we calculate the average number of attention heads that point to the entity mentions being predicted at each layer. We conduct this experiment on NRB with the BERT-large model (24 layers with 16 attention heads at each layer) fine-tuned on CoNLL.

At each layer, we average the number of heads which have their highest attention weight (*argmax*) pointing to the entity name.[8] Figure 5 shows the average number of attention heads that point to an entity mention in the BERT-large model fine-tuned without our methods, with the adversarial noise method (*adv*), and with all three methods.

---

[8] We used the weights of the first sub-token since NRB only contains single word entities.
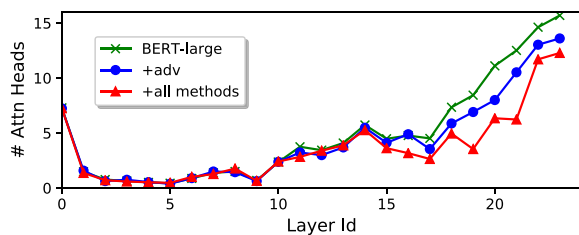
Figure 5: Average number of attention heads (y-axis) pointing to NRB entity mentions at each layer (x-axis) of the BERT-large model fine-tuned on CoNLL.

Figure 6: Performance on NRB of BERT-large models as a function of the number of sentences used to fine-tune them.

Method	π(dev)	π(test)
BERT-large	23.45	25.46
+adv	31.98	31.99
+adv&mask	35.02	34.09
+adv&mask&freeze	**40.39**	**38.62**

Table 4: F1 scores of BERT-large models fine-tuned on CoNLL and evaluated on randomly permuted versions of the dev and test sets: π(dev) and π(test).

We observe an increasing number of heads pointing to entity names when we get closer to the output layer: at the bottom layers (left part of the figure) only a few heads are pointing to entity names, in contrast to the last 2 layers (right part) where almost all heads do so. This observation is in line with Jawahar et al. (2019), who show that bottom and intermediate BERT layers mainly encode lexical and syntactic information, whereas top layers represent task-related information. Our training methods lead to fewer heads at top layers pointing to entity mentions, suggesting the model is focusing more on contextual information.

6.2 Random Permutations

Following the protocol described in Lin et al. (2020), we modified dev and test sets of standard benchmarks by randomly permuting dataset-wise mentions of entities, keeping the types untouched. For instance, the span of a specific mention of a person can be replaced by a span of a location, whenever it appears in the dataset. These randomized tests are highly challenging, as discussed in Section 2, since here the context is the only available clue to solve the task, and many false positive examples are introduced that way.

Table 4 shows the results of the BERT-large model fine-tuned on CoNLL and evaluated on the permuted in-domain dev and test sets. F1
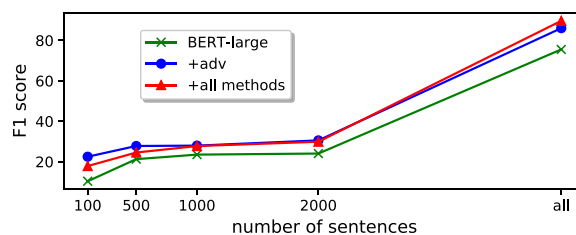
scores are much lower here, confirming this is a hard testbed, but they do provide evidence of the named-regularity bias of BERT. Our training methods improve the model F1 score by 17% and 13% on permuted dev and test sets, respectively, an increase much in line with what we observed on NRB.

6.3 Low Resource Setting

Similarly to Zhou et al. (2019) and Ding et al. (2020), we simulate a low resource setting by randomly sampling tiny subsets of the training data. Since our focus is to measure the contextual learning ability of models, we first selected sentences of CoNLL training data that contain at least one entity followed or preceded by 3 non-entity words.

Then, we randomly sampled $k \in \{100, 500, 1000, 2000\}$ sentences[9] with which we fine-tuned BERT-large. Figure 6 shows the performance of the resulting models on NRB. Expectedly, F1 scores of models fine-tuned with few examples are rather low on NRB as well as on the in-domain test set. Not shown in Figure 6, fine-tuning on 100 and 2000 sentences leads to performance of 14% and 45%, respectively, on the CoNLL test set. Nevertheless, we observe that our training methods, and *adv* in particular, improve performances on NRB even under extremely low resource settings. On CoNLL test and WTS sets, scores vary in a range of ± 0.5 and ± 0.7, respectively, when our methods are added to BERT-large.

6.4 Multilingual Setting

6.4.1 Experimental Protocol

For in-domain data, we use the German, Spanish, and Dutch CoNLL-2002 (Tjong Kim Sang, 2002) NER datasets. Those benchmarks—also from the news domain—come with a train/dev/test split, and the training material is comparable in size

---

[9] $\{0.7, 3.5, 7.1, 14.3\}$% of the training sentences.

	NRB	WTS		NRB	WTS
de	37%	44%	fi	53%	62%
es	20%	22%	da	19%	24%
nl	20%	24%	hr	39%	48%
			af	26%	32%

Table 5: Percentage of translated sentences from NRB and WTS discarded for each language.

to the English CoNLL dataset. In addition, we experiment with four non CoNLL benchmarks: Finnish (Luoma et al., 2020), Danish (Hvingelby et al., 2020), Croatian (Ljubešić et al., 2018), and Afrikaans (Eiselen, 2016) data. These corpora have more diversified text genres, yet mainly follow the CoNLL annotation scheme.[10] Finnish and Afrikaans datasets have comparable size to English CoNLL, Danish is 60% smaller, while the Croatian is twice larger. We use the provided train/dev/test splits for Danish and Finnish, and we randomly split (80/10/10) the Croatian and Afrikaans datasets.

Because NRB and WTS are in English, we designed a simple yet generic method for projecting them to another language. First, both test sets are translated to the target language using an online translation service. In order to ensure a high quality corpus, we eliminate a sentence if the BLEU score (Papineni et al., 2002) between the original (English) sentence and the back translated one is below 0.65.

Table 5 reports the percentage of discarded sentences for each language. While for the Finnish (fi), Croatian (hr), and German (de) languages we remove a large proportion of sentences, we found our translation approach simpler and more systematic than generating an NRB corpus from scratch for each language. The latter approach depends on the robustness of the weak tagger, the number of Wikipedia articles and disambiguation pages per language, as well as the existence of type information. This is left as future work.

For experiments with *fine-tuning*, we use language-specific BERT models[11] for German (Chan et al., 2020), Spanish (Canete et al., 2020),

---

Dutch (de Vries et al., 2019), Finnish (Virtanen et al., 2019), Danish,[12] Croatain (Ulčar and Robnik-Šikonja, 2020), while we use mBERT (Devlin et al., 2019) for Afrikaans.

For *feature-based* approaches, we use the same architecture for ELMo-LSTM (Peters et al., 2018) except that we replace English word embeddings by language-specific ones: FastText (Bojanowski et al., 2017) for static representations, and the aforementioned BERT-base models for contextualized ones.

6.4.2 Results

Table 6 reports the performances on test, NRB, and WTS sets for both *feature-based* and *fine-tuning* approaches with and without our training methods. We used the hyper-parameters of the English CoNLL experiments with no further tuning. We selected the best performing models based on development sets score, and report average results on 5 runs.

Mainly due to implementation details and hyperparameter settings, our fine-tuned BERT-base models perform better on the CoNLL test sets for German (83.8 vs. 80.4) and Dutch (91.8 vs. 90.0) and slightly worse on Spanish (88.0 vs. 88.4) compared to the results reported in their respective BERT papers.

Consistent with the results obtained on English for *feature-based* (Table 1) and *fine-tuned* (Table 3) models, the latter approach performs better on NRB, although by a smaller margin compared to English (+37%). More precisely, we observe a gain of +28% and +26% on German and Croatian respectively, and a gain ranging between 11% and 15% for other languages.

Nevertheless, our training methods lead to systematic and often drastic improvements on NRB coupled with a statistically nonsignificant overall decrease on in-domain test sets. They do, however, incur a slight but significant drop of around 2 F1 score points on WTS for *feature-based* models. Similar to what was previously observed, the best scores on NRB are obtained by BERT models when the training methods are combined. For the Dutch language, we observe that once trained with our methods, the type of models used (feature-based vs. BERT fine-tuned) leads to much less difference on NRB.

Model	German			Spanish			Dutch			Finnish			Danish			Croatian			Afrikaans		
	TEST	NRB	WTS	TEST	NRB	WTS	TEST	NRB	WTS	TEST	NRB	WTS	TEST	NRB	WTS	TEST	NRB	WTS	TEST	NRB	WTS
Feature-based																					
BERT-LSTM	78.9	36.4	84.2	85.6	59.9	90.8	84.9	45.4	85.7	76.0	38.9	84.5	76.4	42.6	78.1	78.0	28.4	79.3	76.2	39.7	65.8
+adv	78.2	44.1	82.8	85.0	65.8	90.2	84.3	57.8	83.5	75.1	52.9	81.0	75.4	47.2	76.9	77.5	35.2	75.5	75.7	42.3	63.3
+adv&mask	78.1	47.6	82.9	84.9	72.2	88.7	84.0	62.8	83.5	74.6	54.3	81.8	75.1	48.4	76.6	76.9	36.8	76.7	75.1	52.8	63.1
Fine-tuning																					
BERT-base	**83.8**	64.0	93.3	**88.0**	72.3	**93.9**	91.8	56.1	**92.0**	**91.3**	64.6	91.9	**83.6**	56.6	86.2	**89.7**	54.7	**95.6**	**80.4**	54.3	91.6
+adv	83.7	68.9	93.6	87.9	75.9	93.9	**91.9**	58.3	91.8	90.2	66.4	92.5	82.7	58.4	**86.5**	89.5	57.9	95.5	79.7	60.2	**92.1**
+a&m&f	83.2	**73.3**	**94.0**	87.4	**81.6**	93.7	91.2	**63.6**	91.0	89.8	**67.4**	92.7	82.3	**63.1**	85.4	88.8	**59.6**	94.9	79.4	**64.2**	91.6

Table 6: Mention level F1 scores of 7 multilingual models trained on their respective training data, and tested on their respective in-domain test, NRB, and WTS sets.

Altogether, these results demonstrate that name regularity bias is not specific to a particular language, even if its degree of severity varies from one language to another, and that the training methods proposed notably mitigate this bias.

7 Conclusion

In this work, we focused on the name regularity bias of NER models, a problem first discussed in Lin et al. (2020). We propose NRB, a benchmark we specifically designed to diagnose such a bias. As opposed to existing strategies devised to measure it, NRB is composed of real sentences with easy to identify mentions.

We show that current state-of-the-art models, perform from poorly (feature-based) to decently (fined-tuned BERT) on NRB. In order to mitigate this bias, we propose a novel adversarial training method based on adding some learnable noise vectors to entity words. These learnable vectors encourage the model to better incorporate contextual information. We demonstrate that this approach greatly improves the contextual ability of existing models, and that it can be combined with other training methods we proposed. Significant gains are observed in both low-resource and multilingual settings. To foster research on NER robustness, we encourage others to report results on NRB and WTS.[13]

This study opens up new avenues of investigation. Conducting a large-scaled multilingual experiment, characterizing the name regularity bias of more diversified morphological language families is one of them, possibly leveraging massively multilingual resources such as WikiAnn (Pan et al., 2017), Polyglot-NER (Al-Rfou et al.,

2015), or Universal Dependencies (Nivre et al., 2016). We can also develop a more challenging NRB by selecting sentences with multi-word entities.

Also, non-sequential labeling approaches for NER like the ones of Li et al. (2020) and Yu et al. (2020) have reported impressive results on both flat and nested NER. We plan to measure their bias on NRB and study the benefits of applying our training methods to those approaches. Finally, we want to investigate whether our adversarial training method can be successfully applied to other NLP tasks.

Acknowledgments

References

Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. 2020a. Entity-switched datasets: an approach to auditing the in-domain robustness of named entity recognition models. *arXiv preprint arXiv:2004.04123*.

Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. 2020b. Interpretability analysis for named entity recognition to understand system predictions and how they can improve. *arXiv preprint arXiv:2004.04564*. **DOI:** `https://doi.org/10.1162/coli_a _00397`

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

---

[13]English and multilingual NRB and WTS are available at `http://rali.iro.umontreal.ca/rali/?q=en /wikipedia-nrb-ner`.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM. **DOI:** https://doi.org/10.1137/1.9781611974010.66

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83. **DOI:** https://doi.org/10.1016/j.csl.2017.01.012

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What's in a name? Are BERT named entity representations just as good for any other name? *arXiv preprint arXiv:2007.06897*. **DOI:** https://doi.org/10.18653/v1/2020.repl4nlp-1.24

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836. **DOI:** https://doi.org/10.18653/v1/D18-1307

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (SEM 2019)*, pages 256–262. **DOI:** https://doi.org/10.18653/v1/S19-1028

Gabriel Bernier-Colborne and Phillippe Langlais. 2020. HardEval: Focusing on challenging tokens to assess robustness of NER. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1697–1704, Marseille, France. European Language Resources Association.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. **DOI:** https://doi.org/10.1162/tacl_a_00051

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. **DOI:** https://doi.org/10.1145/1376616.1376746

Lasse Borgholt, Jakob D. Havtorn, Anders Søgaard Zeljko Agic, Lars Maaløe, and Christian Igel. 2020. Do end-to-end speech recognition models care about context? In *Proceedings of Interspeech*.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020. **DOI:** https://doi.org/10.21437/Interspeech.2020-1750

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *arXiv preprint arXiv:2010.10906*. **DOI:** https://doi.org/10.18653/v1/2020.coling-main.598

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333. **DOI:** https://doi.org/10.18653/v1/P19-1425

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Dont 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 4060–4073.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. *arXiv preprint arXiv:2011.03856*.

Paul R. Cohen. 1996. Empirical methods for artificial intelligence. *IEEE Intelligent Systems*.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Erenay Dayanik and Sebastian Padó. 2020. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391. **DOI:** https://doi.org/10.18653/v1/2020 .acl-main.404

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: A Dutch roberta-based language model. *arXiv preprint arXiv:2001 .06286*. **DOI:** https://doi.org/10.18653 /v1/2020.findings-emnlp.292

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.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. *arXiv preprint arXiv:2011.01549*. **DOI:** https://doi.org/10.18653/v1/2020 .emnlp-main.488

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490. **DOI:** https://doi.org/10.1162/tacl_a _00197

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. **DOI:** https://doi.org/10.18653/v1 /P18-2006

Roald Eiselen. 2016. Government domain named entity recognition for south african languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2016. Robustness of classifiers: from adversarial to random noise. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1632–1640.

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. **DOI:** https://doi.org/10.18653/v1/D19 -1107

Abbas Ghaddar and Phillippe Langlais. 2016. Coreference in Wikipedia: Main concept resolution. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 229–238. **DOI:** `https://doi.org/10.18653/v1/K16-1023`

Abbas Ghaddar and Phillippe Langlais. 2017. Winer: A Wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422.

Abbas Ghaddar and Philippe Langlais. 2018. Transforming Wikipedia into a large-scale fine-grained entity type corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

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. **DOI:** `https://doi.org/10.18653/v1/N18-2017`

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *EMNLP-IJCNLP 2019*, page 132. **DOI:** `https://doi.org/10.18653/v1/D19-6115`

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197. Edinburgh, Scotland, United Kingdom.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*. **DOI:** `https://doi.org/10.18653/v1/2020.acl-main.244`

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780. **DOI:** `https://doi.org/10.1162/neco.1997.9.8.1735`, **PMID:** 9377276

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. Dane: A named entity resource for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4597–4604.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657. **DOI:** `https://doi.org/10.18653/v1/P19-1356`

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. **DOI:** `https://doi.org/10.1162/tacl_a_00300`

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for French. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490.

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

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020.

600

A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859.

Hongyu Lin, Yaojie Lu, Jialong Tang, Xianpei Han, Le Sun, Zhicheng Wei, and Nicholas Jing Yuan. 2020. A rigorous study on named entity recognition: Can fine-tuning pretrained model lead to the promised land? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7291–7300.

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*.

Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018. Training corpus hr500k 1.0. Slovenian language resource repository CLARIN.SI.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for finnish named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4615–4624.

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*, pages 8706–8716. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.769

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60. **DOI:** https://doi.org/10.3115/v1/P14-5010

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie,

Djamé Seddah, and Benoît Sagot. 2019. Camembert: A tasty French language model. *arXiv preprint arXiv:1911.03894*. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.645

Stephen Mayhew, Gupta Nitish, and Dan Roth. 2020. Robust named entity recognition with truecasing pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8480–8487. **DOI:** https://doi.org/10.1609/aaai.v34i05.6368

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. 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 6257–6262. **DOI:** https://doi.org/10.18653/v1/D19-1650

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. **DOI:** https://doi.org/10.18653/v1/P19-1334

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Nafise Sadat Moosavi, Marcel de Boer, Prasetya Ajie Utama, and Iryna Gurevych. 2020. Improving robustness by augmenting training sentences with predicate-argument structures. *arXiv preprint arXiv:2010.12510*.

Thong Nguyen, Duy Nguyen, and Pramod Rao. 2020. Adaptive Name Entity Recognition under highly unbalanced data. *arXiv preprint arXiv:2003.10296*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Sebastian Padó, André Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. 2019. Who sides with whom? Towards computational construction of discourse networks for political debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2841–2847.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. **DOI:** https://doi.org/10.3115/1073083.1073135

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. 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 2227–2237. **DOI:** https://doi.org/10.18653/v1/N18-1202

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. **DOI:** https://doi.org/10.18653/v1/S18-2023

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1596374.1596399

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2020. Learning from others' mistakes: Avoiding dataset biases without modeling them. *arXiv preprint arXiv:2012.01300*.

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 3410–3416. **DOI:** https://doi.org/10.18653/v1/D19-1341

Michael L. Seltzer, Dong Yu, and Yongqiang Wang. 2013. An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Sspeech and Signal Processing*, pages 7398–7402. IEEE. **DOI:** https://doi.org/10.1109/ICASSP.2013.6639100

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.

Anders Søgaard. 2013. Part-of-speech tagging with antagonistic adversaries. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–644.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

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*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: A large-scale dataset for fact extraction and verification. 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 809–819. **DOI:** https://doi.org/10.18653/v1/N18-1074

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. **DOI:** https://doi.org/10.3115/1118853.1118877

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics. **DOI:** https://doi.org/10.3115/1119176.1119195

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest BERT and crosloengual BERT: Less is more in multilingual models. *arXiv preprint arXiv:2006.07890*. **DOI:** https://doi.org/10.1007/978-3-030-58323-1_11

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*.

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.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. **DOI:** https://doi.org/10.18653/v1/W18-5446

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding throughv inference. *arXiv preprint arXiv:1704.05426*. **DOI:** https://doi.org/10.18653/v1/N18-1101

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.204

Yadollah Yaghoobzadeh, Remi Tachet, Timothy J. Hazen, and Alessandro Sordoni. 2019.

Robust natural language inference models with example forgetting. *arXiv preprint arXiv:1911.03861.*

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150.*

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded common-sense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104. **DOI:** `https://doi.org/10.18653/v1/D18 -1009`

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280. **DOI:** `https://doi.org/10.18653/v1 /2020.emnlp-main.590`

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.

Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. 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 357–366. **DOI:** `https://doi.org/10.18653/v1 /D19-1034`

Joey Tianyi Zhou, Hao Zhang, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh, and Kenneth Kwok. 2019. Dual adversarial neural transfer for low-resource named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3461–3471.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764.*