Does Your Model Classify Entities Reasonably? Diagnosing and Mitigating Spurious Correlations in Entity Typing

Nan Xu, Fei Wang, Bangzheng Li, Mingtao Dong, Muhao Chen
Department of Computer Science & Information Sciences Institute
University of Southern California
{nanx,fwang598,bangzhen,mingtaod,muhaoche}@usc.edu

Abstract

Entity typing aims at predicting one or more words that describe the type(s) of a specific mention in a sentence. Due to shortcuts from surface patterns to annotated entity labels and biased training, existing entity typing models are subject to the problem of *spurious correlations*. To comprehensively investigate the faithfulness and reliability of entity typing methods, we first systematically define distinct kinds of model biases that are reflected mainly from spurious correlations. Particularly, we identify six types of existing model biases, including *mention-context* bias, *lexical overlapping* bias, *named entity* bias, *pronoun* bias, *dependency* bias, and *overgeneralization* bias. To mitigate model biases, we then introduce a counterfactual data augmentation method. By augmenting the original training set with their debiased counterparts, models are forced to fully comprehend sentences and discover the fundamental cues for entity typing, rather than relying on spurious correlations for shortcuts. Experimental results on the *UFET* dataset show our counterfactual data augmentation approach helps improve generalization of different entity typing models with consistently better performance on both the original and debiased test sets[1].

1 Introduction

Given a sentence with an entity mention, the *entity typing* task aims at predicting one or more words or phrases that describe the type(s) of that specific mention (Ling and Weld, 2012; Gillick et al., 2014; Choi et al., 2018). This task essentially supports the structural perception of unstructured text (Distiawan et al., 2019), being an important step for natural language understanding (NLU). More specifically, entity typing has a broad impact on various NLP tasks that depend on type understanding, including coreference resolution (Onoe and Durrett, 2020), entity linking (Hou et al., 2020; Tianran et al., 2021), entity disambiguation (Onoe and Durrett, 2020), event detection (Le and Nguyen, 2021) and relation extraction (Zhou and Chen, 2022).

To tackle the task, literature has developed vari-

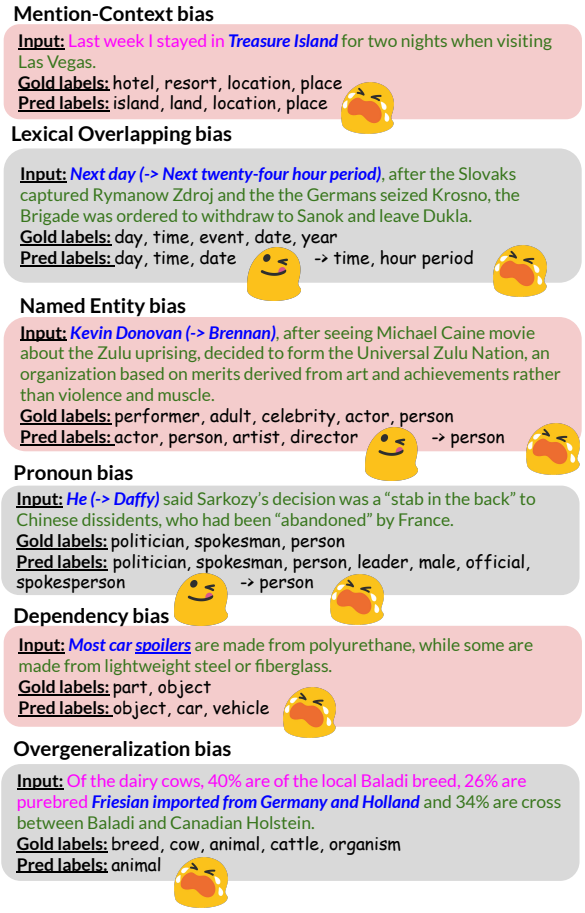

Figure 1: Examples demonstrating spurious correlations exploited by one of the SOTA entity typing models *MLMET*. Left context is in magenta, *entity mention* in italic blue, right context in green. Perturbations upon mentions and new predictions start from →/ →. implies good predictions by exploiting spurious correlations and indicates bad predictions when spurious correlations no longer exist. *MLMET* falsely relies on the entity name to give "island" predictions for a hotel mention, incorrectly infers types of the dependent "car" rather than the headword "spoiler", and gives only the coarse label "animal" with more fine-grained missing.

[1]Code and resources are available at https://github.com/luka-group/DiagnoseET.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8642 - 8658
December 7-11, 2022

ous predictive methods to capture the association between the contextualized entity mention representation and the type label. For instance, a number of prior studies approach the problem as multi-class classification based on distinct ways of representing the entity mentioning sentences (Yogatama et al., 2015; Ren et al., 2016; Xu and Barbosa, 2018; Dai et al., 2021). Other studies formulate the problem as structured prediction and leverage structural representations such as box embeddings (Onoe et al., 2021) and causal chains (Liu et al., 2021) to model the dependency of type labels. However, due to shortcuts from surface patterns to annotated entity labels and biased training, existing entity typing models are subject to the problem of *spurious correlation* (Wang and Culotta, 2020; Wang et al., 2021; Branco et al., 2021). For example, given a sentence "Last week I stayed in *Treasure Island* for two nights when visiting Las Vegas.", a SOTA model like *MLMET* (Dai et al., 2021) may overly rely on the entity name and falsely type *Treasure Island* as an *island*, while ignoring the sentential context that indicates this entity as a *resort* or a *hotel*. For morphologically rich mentions with multiple noun words such as "most car spoilers", entity models may fail to understand its syntactic structure and miss the target entity from the actual head-dependent relationship, leading to predictions describing the dependent car (*car*, *vehicle*) rather than the head spoilers (*part*). Such spurious clues can cause the models to give unfaithful entity typing and further harm the machine's understanding of the entity mentioning text.

To comprehensively investigate the faithfulness and reliability of entity typing methods, the *first* contribution of this paper is to systematically define distinct kinds of model biases that are reflected mainly from spurious correlations. Particularly, we identify the following six types of existing model biases, for which examples are illustrated in Fig. 1. Those biases include *mention-context* biases, *lexical overlapping* biases, *named entity* biases, *pronoun* biases, *dependency structure* biases and *overgeneralization* biases. We provide a prompt-based method to identify instances posing those biases to the typing model. In the meantime, we illustrate that common existence of these types of biased instances causes it hard to evaluate whether a model is faithfully comprehending the entire context to infer the type, or trivially leveraging surface forms or distributional cues to guess the type.

We introduce a counterfactual data augmentation (Zmigrod et al., 2019) method for debiasing entity typing, as the *second* contribution of this paper. Given biased features, we reformulate entity typing as a type-querying cloze test and leverage a pre-trained language model (PLM) to fill in the blank. By augmenting the original training set with their debiased counterparts, models are forced to fully comprehend the sentence and discover the fundamental cues for entity typing, rather than rely on spurious correlations for shortcuts. Compared with existing debiasing approaches such as product of experts (He et al., 2019), focal loss (Karimi Mahabadi et al., 2020), contrastive learning (Zhou et al., 2021) and counterfactual inference (Qian et al., 2021), our counterfactual data augmentation approach helps improve generalization of all studied models with consistently better performance on both original *UFET* (Choi et al., 2018) and debiased test sets.

2 Method

In this section, we start with the problem definition (§2.1) and then categorize and diagnose the spurious correlations causing shortcut predictions by the typing model (§2.2). Lastly, we propose a counterfactual data augmentation approach to mitigate the identified spurious correlations, as well as several alternative techniques that apply (§2.3).

2.1 Problem Definition

Given a sentence s with an entity mention $e \in s$, the *entity typing* task aims at predicting one or more words or phrases T from the label space L that describe the type(s) of e.

By nature, the inference of type T should be context-dependent. Take the first sample demonstrated in Fig. 1 as an instance: in "Last week I stayed in ***Treasure Island*** for two nights when visiting Las Vegas," Treasure Island should be typed as *hotel* and *resort*, rather than *island* or *land* by trivially considering the surface of mention phrase.

2.2 Spurious Correlations Diagnoses

We systematically define six types of typical model biases caused by spurious correlations in entity typing models. For each bias, we qualitatively inspect its existence and the corresponding spurious correlations used by a SOTA entity typing model on sampled instances with bias features. Following Poerner et al. (2020), we prompt a PLM, *RoBERTa-*

PLM Prompts	Entity Typing Instances
Mention-Context: **Prompt I**: <*Mention*> *is a type of* <*mask*>. **S1**: ***fire*** is a type of <*mask*>. ***RoBERTa***: *energy, heat, explosion, fire, gas* **S2**: ***the war*** is a type of <*mask*>. **True labels**: *war, battle, conflict* ***RoBERTa***: *war, battle, conflict, violence, warfare*	**T1**: A teacher who survived the shooting said he would never forgive the police for taking an hour to arrive after the gunman opened ***fire***. **True labels**: *injury, shooting, event, violence* ***MLMET***: *event, object, fire* (F1: 0.285) **T2**: ***fire*** ***MLMET***: *object, fire* (F1: 0.0)
Lexical Overlapping: **Prompt II**: <*Left Context*> <*New Mention with Word Substitution*> <*Right Context*>. <*New Mention with Word Substitution*> *is a type of* <*mask*>. **S3**: ***Next twenty-four hour period***, after the Slovaks captured ... ***Next twenty-four hour period*** is a type of <*mask*>. ***RoBERTa***: *confusion, retreat, ambush, battle, timeline*	**T3**: ***Next day***, after the Slovaks captured Zdroj the Brigade was ordered to withdraw to Sanok. **True labels**: *day, time, event, date, year* ***MLMET***: *day, time, date* (F1: 0.749) **T4**: ***Next twenty-four hour period***, after the Slovaks captured Zdroj, the Brigade was ordered to withdraw to Sanok. ***MLMET***: *time, hour, period* (F1: 0.25)
Named Entity: **Prompt III**: *The* <*Attribute*> <*Named Entity*> *is a type of* <*mask*>. **S4**: The person ***Benjamin Netanyahu*** is a type of <*mask*>. ***RoBERTa***: *politician, person, character, personality, man* **S5**: The person ***Jintara Poonlarp*** is a type of <*mask*>. ***RoBERTa***: *person, human, woman, personality, character*	**T5**: ***Benjamin Netanyahu*** asserted that Amin al-Husseini had been one of the masterminds of the Holocaust. **True labels**: *politician, leader, person* ***MLMET***: *politician, leader, person* (F1: 1.0) **T6**: ***Jintara Poonlarp*** asserted that Amin al-Husseini had been one of the masterminds of the Holocaust . ***MLMET***: *person, scholar, writer* (F1: 0.333)
Pronoun: **Prompt IV**: <*Left Context*> <*Person Name*> <*Right Context*>. <*Person Name*> *is a type of* <*mask*>. **S6**: ***Judith*** other film credits include " The Four Feathers," " Dr. T & the Women " and " 200 Cigarettes." ***Judith*** is a type of <*mask*>. ***RoBERTa***: *bird, cat, vampire, rabbit, dog*	**T7**: ***Her*** other film credits include " The Four Feathers," " Dr. T & the Women " and " 200 Cigarettes." **Truths**: *woman, performer, adult, female, entertainer, person, actress* ***MLMET***: *woman, female, actress, person, artist* (F1: 0.666) **T8**: ***Judith*** other film credits include " The Four Feathers," " Dr. T & the Women " and " 200 Cigarettes." ***MLMET***: *person* (F1: 0.25)

Table 1: Entity typing instances with content-based biases recognized by RoBERTa-large (Liu et al., 2019). To reflect the shortcuts exploited by entity typing models (§2.2), we list the sentences, labels and predictions from one of the SOTA models *MLMET* in T1, T2, T3, T5 and T7. To identify biased instances (§2.2), we show the constructed masked fill-in task to query the PLM with mention types from S1 to S6. To mitigate spurious correlations (§2.3), we show the proposed counterfactual data augmentation where the shortcuts disappear and the model fails in T4, T6 and T8. We underline the ***mention span*** in italic boldface and record the macro F1 score for each prediction.

large (Liu et al., 2019), to identify potential biasing samples with either detected surface patterns or facts captured during training. To do so, we reformulate entity typing as a type-querying cloze task and perform the analysis as follows.

1) Mention-Context Bias: Semantically rich entity mentions may encourage the model to overly associate the mention surface with the type without considering the key information stated in contexts. An example is accordingly shown in T1 of Tab. 1, where *MLMET* predicts types that correspond to the case where "fire" is regarded as *burning* instead of *gun shooting*. Evidently, this is due to not effectively capturing the clues in the context such as "shooting" and "gunman". This is further illustrated by the counterfactual example T2, where the model predicts almost the same labels when seeing "fire" without a context.

To identify potential instances with the mention-context bias, we query the PLM to infer the entity types based only on the mention with the template shown in Prompt I (Tab. 1). Therefore, samples where the PLM can accurately predict without the context information are regarded as biased. Entity typing models can easily achieve good performance on those biased samples by leveraging spurious correlations between their mention surface and types, as shown in S2 from Tab. 1.

2) Lexical Overlapping Bias: Type labels that have lexical overlaps with the entity mention can also become prediction shortcuts. As shown in **T3** from Tab. 1: labeling mention "next day" with the type *day* and additional relevant types leads to the F1 up to 0.749. We observe a considerable amount of similar examples, e.g., typing the mention "eye shields" as *shield*, "the Doha negotiations" as *negotiation*, etc. The highly overlapped mention words and type labels make it difficult to evaluate whether the model makes predictions based on content comprehension or simply lexical similarities.

Bias Type	Analyses	Entity Typing Instances
Dependency: Models fail to capture the syntactic structure and make type predictions focusing on other components rather than the dependency headword	**S6 Dependency Parsing**: token | text | head text 1st whale anatomy 2nd anatomy anatomy	**T9**: Dubois contributed an article on ***whale anatomy*** to a book by the Dutch zoologist , Max Wsubjecteber , and , inspired by the fresh discovery of new Neanderthal fossils at the Belgian town of Spy , he spent his vacation fossil hunting in the vicinity of his birthplace . **True labels**: *subject, topic* ***MLMET***: *object, animal* (F1: .0) **T10**: Dubois contributed an article on ***anatomy*** to a book by the Dutch zoologist , Max Wsubjecteber , and , inspired by the fresh discovery of new Neanderthal fossils at the Belgian town of Spy, he spent his vacation fossil hunting in the vicinity of his birthplace . ***MLMET***: *object, concept, subject* (F1: .4)
Overgeneralization: Models suffer from biased training due to extreme class imbalance	**Statistics**: **Counts of Coarse Types**: *person*: 824, *event*: 181 **Counts of Ultra-fine Types**: *concept*: 68, *activity*: 23, *trouble*: 8, *difficulty*: 6, *problem*: 5, *misconduct*: 1, *use*: 1, *abuse*: 0 *behavior*: 0, *wrongdoing*: 0	**T11**: ***Many nineteenth-century individualist anarchists ,*** including Benjamin Tucker , rejected the anarcho-capitalist Lockean position in favour of the anarchist position of " occupancy and use " -LRB- or " possession " , to use Proudhon 's term -RRB- , particularly in land . **True labels**: *person* ***MLMET***: *person* (F1: 1.0) **T12**: In a letter , the exchange said its investigations had turned up " no evidence of ***abusive behavior*** . " **True labels**: *behavior, wrongdoing, difficulty, misconduct, trouble, use, activity, problem, concept, abuse* ***MLMET***: *behavior, event* (F1: .166) **T13**: <NULL Input> ***MLMET***:*person* (prob=0.992)

Table 2: Entity typing instances from *UFET* test set with biases detected based on statistical analyses. To discover shortcuts utilized by entity typing models (§2.2), we show one *Dependency* bias instance where the model fails to locate the target entity in the mention (T9) and two *Overgeneralization* bias instances: T11 annotated by coarse types and T12 annotated by ultra-fine types. To quantify the overgeneralization bias (§2.2), we query the typing model with an empty sentence in T13. To mitigate spurious correlations (§2.3), we do dependency parsing to distinguish headwords from dependents in S6 and truncate the mention with only the headword preserved as T10 to help address dependency bias.

We substitute the overlapping mention words with semantically similar words and ask the PLM to infer the entity types on such perturbed instances (details introduced in §2.3) by prompting with the template Prompt II (Tab. 1). We consider instances have lexical overlapping biases when the PLM performs poorly after the overlapped mention words are substituted, as shown in S3 of Tab. 1.

3) Named Entity Bias: On cases where mentions refer to high-reporting entities in corpora, models may be trained to ignore the context but directly predict labels that co-occur frequently with those entities. We show a concrete instance to type a person named entity in **T5** of Tab. 1. The mention *Benjamin Netanyahu*, known as Israeli former prime minister, is normally annotated with *politician, leader* and *authority*. After observing popular named entities and their common annotations during training, models are able to predict their common types, making it hard to evaluate models' capabilities to infer context-sensitive labels.

As illustrated in Prompt III (Tab. 1), we prompt the PLM to type the named entity when only the name and its general attribute is given, e.g., *the geopolitical area India* or *the organization Apple*, etc. We regard instances to have the named entity bias when the PLM accurately infers the mention types relying on prior knowledge of named entities. In Tab. 1, we show one instance with the mention containing *Benjamin Netanyahu* in S4, and the Thai pop music singer – *Jintara Poonlarp* in S5[1]. Based on types related to *Benjamin*'s political role in S4 and general types for *Jintara* in S5, we consider instances to type mentions including *Benjamin* as biased while those with *Jintara* as unbiased.

4) Pronoun Bias: Compared with diverse person names, pronouns show up much more frequently to help make sentences smoother and clearer. Therefore, models are subject to biased training to type pronouns well, but lose the ability to type based on diverse real names. To type the pronoun *her* in **T7** of Tab. 1, the entity typing model can successfully infer general types *woman, female* as well as the context-sensitive type *actress*. To obtain high generalization, we expect models to infer types correctly for both pronouns and their referred names.

We substitute the gender pronoun with a random person name of the same gender (details introduced in §2.3) and ask the PLM to infer the types with Prompt IV (Tab. 1). We consider samples to have the pronoun bias when the PLM fails to capture the majority of types after the name substitution, as

[1]Both represent celebrities reported by their own Wikipedia pages and thousands of news articles, hence are very likely to be covered by the pre-training corpora of *RoBERTa*.

shown in S6 of Tab. 1.

5) Dependency Bias: It is observed that the mention's headwords explicitly match the mention to its types (Choi et al., 2018). However, models may fail to capture the syntactic structure with predictions focusing on dependents instead of headwords. We show an instance with inappropriate focus among mention words in **T9** of Tab. 2.Without understanding the mention's syntactic structure, entity typing models may make predictions that are irrelevant to the actual entity.

Since knowledge about mention structures is beneficial for typing complex multi-word mentions, we mitigate the bias by data augmentation to improve model learning (details introduced in §2.3), rather than identify whether the bias exists or not.

6) Overgeneralization Bias: When training with disproportional distributed labels, frequent labels are more likely to be predicted compared with rare ones. Entity typing datasets are naturally imbalanced (Gillick et al., 2014; Choi et al., 2018). We show two instances annotated by coarse- and fine-grained labels in **T11** and **T12** of Tab. 2: the model can easily predict the coarse-grained label *person* to describe "anarchist", but fails to infer less frequent but more concrete labels such as *misconduct* and *wrongdoing* to type *behavior*. Models ought to type entities by reasoning on mentions and contexts, rather than trivially fitting the label distribution.

As shown in T13 of Tab. 2, we craft a special instance – an empty sentence, with which the uniform distribution over all types is expected from models free of overgeneralization bias. We then compute its disparity with the model's actual probability distribution: the higher/lower probability predicted on popular/rare types, the more biased the model on the label distribution.

Discussion The prior defined six biases are not mutually exclusive. We discuss some possible mixtures of concurrent biases as follows:

Mention-Context and *Lexical Overlapping Bias*: the model falsely types the mention "Treasure Island" as *island*, without understanding the context talking about the holiday accommodation. Another possible reason that the mention far outweighs the context might be the high word similarity between mention word "Island" and type word "island".

Dependency and *Lexical Overlapping Bias*: *MLMET* incorrectly makes the prediction *car* for the mention "most car spoilers" without distinguishing important headwords from less important dependent words. Another reasonable explanation for emphasizing on the dependent rather than the headword is its perfect lexical match with the type set, where "car" is a relatively popular label but no type has high word similarity with "spoilers". To diagnose and mitigate all spurious correlations the entity typing model may take advantage of, we disentangle the multiple biases on a single instance by *analyzing each bias individually without considering their mutual interactions*.

2.3 Mitigating Spurious Correlations

Models exploiting spurious correlations lack the required reasoning capability, leading to unfaithful typing and harmed out-of-distribution generalization when bias features observed during training do not hold. Therefore, we propose to mitigate spurious correlations from the counterfactual data augmentation perspective: for each instance recognized with specific bias features, we automatically craft its debiased counterpart and train entity typing models with both samples. Whenever the model prefers to exploit biasing features, it will fail on newly crafted debiased instances and actively look for more robust features: understanding and reasoning on the sentence rather than exploiting spurious correlations. Considering the characteristic textual patterns from different biases, we propose the following distinct strategies to craft debiased instances for four types of biases (with examples explained in Appx. §A.1). Note that although we can hardly craft a new instance free of mention-context bias or overgeneralization bias, we can choose to leverage the alternative debiasing techniques introduced in later parts of this section for these two biases.

Counterfactual Augmentation On instances diagnosed with lexical overlapping biases, we perform word substitutions in two steps to substitute mention words lexically similar to type labels with original semantics preserved. To do so, we identify the sense of type words in mentions using an off-the-shelf word sense disambiguation model (Barba et al., 2021) and substitute them with their WordNet synonyms. We consider perturbed sentences with poor performance from the PLM as the *counterfactual augmented* instances free from lexical overlapping bias, to prohibit the entity typing model from exploiting spurious correlations (T4 of Tab. 1).

For instances with the named entity bias, we augment by performing named entity substitution according to the following criteria. *1) validity*: sub-

stituted entities should have the same general type as the original ones[1], e.g., the geopolitical area "India" can be replaced by "London"; *2) debiased*: models training on large corpora should not possess comprehensive knowledge of the new named entities. Basically, we leverage an off-the-shelf NER model (Ushio and Camacho-Collados, 2021) to identify and classify named entities into general NER types provided by this model, and then divide the entities into **informative** and **non-informative** group based on the prompt-based typing performance by the PLM. We then substitute informative named entities with non-informative ones sharing the same NER type as the *counterfactual augmented* instances (T6 of Tab. 1).

For the pronoun bias, we craft new instances by concretizing pronoun mentions in two situations. If co-reference resolution (Toshniwal et al., 2021) detects the referred entity of the pronoun mention in the context, that entity is selected as the new mention. Otherwise, the gender pronoun mention will be substituted with a randomly sampled masculine/feminine name from the NLTK corpus (Bird, 2006). New sentences with the actual person names are considered *counterfactual augmented* if the PLM fails to infer the person's type with contextual information given (T7 of Tab. 1).

We further augment from instances where mentions have internal dependency structures to tackle the dependency bias. First, we use a dependency parsing tool (Honnibal et al., 2020) to recognize the dependency parse tree of the mention. On top of that, we truncate all other dependent words in the new mention to create the augmentation. From associations between explicitly provided headwords and their matching labels, the models are encouraged to learn dependency structures for targeted entity typing and predict precisely when headwords and dependents are mixed in mentions (T9 of Tab. 2).

Together with the new instances with headwords explicitly given, instances counterfactually augmented upon the entity typing training set is utilized to allow various entity typing models to learn to mitigate spurious correlations. Meanwhile, we leverage the counterfactual augmented instances derived from the test set for model evaluation.

Alternative Debiasing Techniques In addition to data augmentation, other applicable debiasing techniques can be used to resample or reweight original instances in training, or directly measure and deduct biases in inference. A typical resampling technique is *AFLite* (Le Bras et al., 2020) which drops samples predicted accurately by simple models such as fasttext (Joulin et al., 2017). Reweighting techniques typically train one or more models to proactively identify and upweight underrepresented instances in the training process, which includes product of experts, debiased focal loss, learned-mixin and its variant learned-mixin+H (Clark et al., 2019; He et al., 2019; Karimi Mahabadi et al., 2020). On the other hand, counterfactual inference (Qian et al., 2021) measures prediction biases based on counterfactual examples (e.g. masking out the context for measuring mention-context biases, or giving empty inputs to measure overgeneralization biases (Wang et al., 2022)), and directly deducts the biases in inference. In addition, contrastive learning (Chen and He, 2021; Caron et al., 2021; Chen and He, 2021) can be used to adopt a contrastive training loss (Caron et al., 2021; Chen and He, 2021) to discourage the model from learning similar representations for full and bias features[2]. Next, we compare our approach with those techniques.

3 Experiments

In this section, we start with describing the experimental setups (§3.1). Next, we diagnose entity models to measure their reliance on spurious correlations (§3.2). We then compare our counterfactual data augmentation with other debiasing techniques for spurious correlation mitigation (§3.3).

3.1 Experimental Settings

We leverage the ultra-fine entity typing (**UFET**) dataset (Choi et al., 2018) to evaluate entity typing models and apply different mitigation approaches either during training or as inference post-processing. UFET comes with $6K$ samples from crowdsourcing and $25.2M$ distant supervision samples. There are $10,331$ types in total, among which nine are general (e.g., person), 121 are fine-grained (e.g., engineer), and $10,201$ are

[1]We consider the 12 NER types including person, geopolitical area, location, organization, group, date, facility, work of art, ordinal number, event, product, and time.

[2]Models are discouraged to learn similar representations between full features and bias features such as mention, named entity, or pronoun as input. They are also encouraged to learn similar representations between original instances and counterfactual augmented instances with overlapped word substitution, instances with headwords as new mentions.

Test Set	Mention-Context 131/1085 (↑)		Lexical Overlapping 475/3 (↓)		Named Entity 36/500 (↓)		Pronoun 881/20 (↓)		Dependency 280, 222/961		Overgeneralization 93/242	
	BiLSTM	MLMET	BiLSTM	MLMET	BiLSTM	MLMET	BiLSTM	MLMET	BiLSTM	MLMET	BiLSTM	MLMET
Biased	.385	.654	.510	.551	.504	.735	.494	.561	.152	.424	.466	.427
-Perturb.	.400 (3.8%)	.654 (0.0%)	.050 (-90.3%)	.327 (-40.8%)	.332 (-34.0%)	.600 (-18.4%)	.179 (-63.8%)	.525 (-6.4%)	.396 (160.5%)	.564 (32.9%)	.118	.232
Unbiased	.265	.436	.392	.544	.316	.505	.366	.683	-	-	-	-
-Perturb.	.253 (-4.7%)	.395 (-9.4%)	.167 (-57.5%)	.444 (-18.2%)	.372 (17.6%)	.539 (6.9%)	.130 (-64.6%)	.659 (-3.5%)	-	-	-	-

Table 3: F1 scores of two representative entity typing models on UFET testing samples with(out) distinct biases and their perturbations: mention-only input for *Mention-Context*, overlapped word substitution for *Lexical Overlapping*, named entity substitution for *Named Entity*, name substitution for *Pronoun*. Below each bias, the number of model-agnostic biased and unbiased instances are listed and ↓ / ↑ indicates expected performance from models leveraging spurious correlations after perturbing biased instances. Relative performance drop/increase after testing on their perturbations is recorded in brackets. For *Dependency* bias, we show performance on 280 and 222 out of 961 test samples where the two models benefit from making predictions based on headwords and contexts respectively. For *Overgeneralization* bias, we show performance on 93/242 samples annotated by purely coarse/ultra-fine types (values on different subsets hence incomparable). See results of all five models evaluated by full metrics in Tab. 7.

ultra-fine (e.g., flight engineer). We follow prior studies (Choi et al., 2018) to evaluate entity typing models with macro-averaged precision, recall and F1. We also study spurious correlations and effectiveness of the proposed debiasing approach on **OntoNotes** (Gillick et al., 2014). As results present similar observations, we leave detailed analysis in Appx. §A.3.

Entity Typing Baselines We diagnose the prediction biases and the effectiveness of distinct debiasing models based on following approaches: 1) **BiLSTM** (Choi et al., 2018) concatenates the context representation learned by a bidirectional LSTM and the mention representation learned by a CNN to predict entity labels. 2) **LabelGCN** (Xiong et al., 2019) introduces graph propagation to encode global label co-occurrence statistics and their word-level similarities. 3) **LRN** (Liu et al., 2021) autoregressively generates entity labels from coarse to fine levels, modeling the coarse-to-fine label dependency as causal chains. 4) **Box4Types** (Onoe et al., 2021) proposes to embed concepts as d-dimensional hyper rectangles (boxes), so that hierarchies of types could be captured as topological relations of boxes. 5) **MLMET** (Dai et al., 2021) augments training data by constructing mention-based input for BERT to predict context-dependent mention hypernyms for type labels. Without loss of generality, we discuss results of two representative models, the earliest *BiLSTM* training from scratch and the latest *MLMET* finetuning on the PLM, for the sake of clarity in this section. As the observations on the other models are similar, we leave those results in Appx. §A.

3.2 Diagnosing Entity Typing Models

In Tab. 3, we report performance of entity typing models trained on UFET. The models are tested on original biased samples and their perturbed new instances to reflect exploited spurious correlations. We conduct similar analyses on unbiased samples.

1) Mention-Context Bias: When perturbing the biased samples by only feeding their mentions to typing models, the performance of *MLMET* keeps unchanged while the performance of *BiLSTM* even improves by 3.8%. This disobeys the task goal of entity typing where types of the mentions should also depend on contexts, and we suggest that samples with mention-context biases are insufficient for a faithful evaluation of a reliable typing system.

2) Lexical Overlapping Bias: After substituting label-overlapped mention words with semantically similar words, performance of both models drops drastically especially on biased samples identified by the PLM. Compared with *MLMET*, *BiLSTM* has less parameter capacity and is more inclined to leverage lexical overlapping between mentions and type labels as the shortcut for typing.

Compared with original biased instances, the perturbed instances with label-overlapped mention words replaced might look less natural or fluent. In Tab. 4, we therefore substitute words from different parts of instance, and prove that performance degradation is caused by removed lexical overlapping bias rather than unnatural or dysfluent input.

3) Named Entity Bias: After replacing named entities to be less impacted from biased prior knowl-

Perturbed Words	BiLSTM	MLMET
Label-overlapped Mention	-63.84%	-38.60%
Non Label-overlapped Mention	-2.77%	-11.64%
Context	3.86%	-0.18%

Table 4: F1 performance variation after perturbing words at different parts of instances from UFET test with their synonyms. Much higher performance drop after replacing label-overlapped mention words proves the degradation caused by removing label overlapping bias rather than potential reduced naturalness or fluency due to word substitution.

Model	Top/Bottom Types (Prob.)
BiLSTM	person (.928), title (.437), concept (.104)... ...vice squad (.000), adolescent (.000), supporter (.000)
LabelGCN	concept (.349), increase (.249), case (.192)... ...archipelago (.000), spiritual leader (.000), national park (.000)
Box4Types	object (.626), person (.282), company (.231)... ...dismissal (.000), trump (.000), small town (.000)
LRN	person (.998), writer (.000), place (.000)... ...chicken leg (.000), chicken wing (.000), chicken wire (.000)
MLMET	person (.992), time (.314), title (.100)... ...consortium (.000), negotiator (.000), football player (.000)

Table 5: Top and bottom predictions and their probabilities when querying typing models with empty input.

edge, performance of both studied models in Tab. 3 decreases considerably when encountering named entities, with which models struggle to capture spurious correlations with mention types. Interestingly, perturbing unbiased samples by utilizing named entities with bias provides shortcuts for prediction, leading to improved performance of both models.

4) Pronoun Bias: With pronouns replaced by their referred entities in contexts or random masculine/feminine names otherwise, we observe serious performance degradation from both models, which demonstrates their common weakness on typing more diverse and less frequent real names.

5) Dependency Bias: With headwords directly exposed to entity typing models by dropping all other less important dependents, performance from *BiLSTM* on around 30% of all testing samples with dependency structures gets improved dramatically, while *MLMET* also predicts more precisely on 23% of samples. Hereby, we confirm that existing entity models still suffer from extracting core components of given mentions for entity typing and appeal for more research efforts to address this problem.

6) Overgeneralization Bias: Models are subject to making biased predictions towards popular types observed during training, which leads to contrastive performance on instances purely annotated by coarse and ultra-fine types, as shown in Tab. 3. This problem is exemplified in a case study in Tab. 5, where typing models are queried with an empty sentence. Compared with the uniform probability distribution expected from models free from overgeneralization bias, existing models are inclined to give much higher probabilities to coarse types such as *person* and *title*.

3.3 Mitigating Spurious Correlations

In Tab. 6, we evaluate robustness of entity typing models after adopting the proposed counterfactual data augmentation or alternative debiasing techniques, and present results on the *UFET* test set with bias and our counterfactually debiased test set.

Overall, our counterfactual data augmentation is the only approach that consistently improves the generalization of the studied models across both test sets. Particularly, we achieve the best performance on *UFET* and the debiased test set with *MLMET*. Besides, models trained with our approach improve the performance of *BiLSTM* and *MLMET* relatively by 71.15% and 11.81% on the debiased test set, respectively, implying the least reliance on spurious correlations to infer correct entity types.

When evaluating other debiasing approaches, we find that 1) none of the resampling or reweighting techniques is capable to maintain the performance on UFET test set of both models, which could be attributed to the large-scale label space and the existence of diverse causes of model biases; 2) contrastive learning with either cross entropy loss or cosine similarity loss helps improve performance on debiased samples, but leads to accuracy drop of *MLMET* on UFET; 3) without updating model parameters given bias features, counterfactual inference fails to improve performance of *MLMET* on debiased samples.

4 Related Work

Entity Typing Earlier studies on entity typing (Yogatama et al., 2015; Ren et al., 2016; Xu and Barbosa, 2018) learned contextual embeddings for entity mentions and types to capture their association. To model label correlations without annotated label hierarchies in UFET, LabelGCN (Xiong et al., 2019) introduced the graph propagation layer to encode global label co-occurrence statistics and their word-level similarities, whereas HMGCN (Jin et al., 2019) proposed to infer this information from a knowledge base. For the same purpose,

Approach	BiLSTM						MLMET					
	U-Prec.	U-Rec.	U-F1	A-Prec.	A-Rec.	A-F1	U-Prec.	U-Rec.	U-F1	A-Prec.	A-Rec.	A-F1
No Debiasing	.471	.242	.320	.242	.182	.208	**.527**	.452	.487	**.554**	.414	.474
AFLite	***.529***	.209	.300	*.296*	.142	.192	.526	.450	.485	.551	*.420*	*.477*
POE	.441	*.281*	***.343***	*.339*	*.267*	*.299*	.518	*.458*	.486	.543	*.425*	*.477*
Focal	.315	***.341***	*.328*	.207	*.269*	*.234*	.520	*.460*	*.488*	.545	*.419*	.474
Learned-mixin	.396	*.289*	*.335*	*.282*	*.279*	*.280*	.483	*.472*	.478	.480	*.453*	.466
Learned-mixin+H	.279	*.326*	.301	.182	*.338*	*.237*	.405	***.529***	.459	.379	*.485*	.425
Contrastive (CE)	.461	*.272*	*.342*	*.312*	*.354*	*.331*	.477	*.471*	.474	.449	*.460*	.454
Contrastive (Cosine)	.440	*.257*	*.325*	***.462***	*.265*	*.337*	.495	.451	.472	.489	*.441*	.464
Counterfact. Inf.	.442	*.264*	*.331*	*.347*	.173	*.231*	.525	*.454*	.487	.492	*.446*	.468
Augmentation (ours)	*.473*	*.260*	*.336*	*.345*	***.367***	***.356***	.515	*.466*	**.489**	.540	***.470***	***.530***

Table 6: Effectiveness of different debiasing approaches on two representative entity typing models when testing on UFET test set (U-) and our counterfactual augmented test set (A-). The best performance per column is marked in **bold** while improved values over those without debiasing in *italic*. For contrastive learning, CE stands for the cross entropy and Cosine represents cosine similarity. See results of three other entity typing models in Tab. 10.

Box4Types (Onoe et al., 2021) was proposed to embed concepts as hyper rectangles (boxes), such that their topological relations can represent type hierarchies. Considering the prevailing noisy labels in existing entity typing datasets, Onoe and Durrett (2019) performed supervised denoising to filter and fix noisy training labels. Dai et al. (2019) introduced distant supervision from entity linking results. To tackle the sparsity of training, recent work conducted data augmentation with a masked language model and WordNet knowledge to enrich the training data (Dai et al. 2021; MLMET), and made use indirect supervision from natural language inference (Li et al. 2022; LITE). Despite much attention in literature, to the best of our knowledge, our work represents the first investigation on faithfulness and reducing shortcuts in this task.

Spurious Correlations in NLP Models Much recent effort has been put into studying spurious correlation in Natural Language Inference (NLI) tasks. Recent studies show that crowd workers are prone to produce annotation artifacts (Gururangan et al., 2018) through the rapid annotation process and result in identifiable shortcut features (Karimi Mahabadi et al., 2020; Du et al., 2021a). Hence, simple models can easily achieve good performance even with partial inputs (Kaushik et al., 2019; Karimi Mahabadi et al., 2020), or leveraging superficial syntactic properties (McCoy et al., 2019; Utama et al., 2020; Pezeshkpour et al., 2021). On several other NLP tasks composed of multiple textual components, it has been observed that models fed with partial inputs can already achieve competitive performance, e.g., predicting for claim verification (Schuster et al., 2019; Utama et al., 2020; Du et al., 2021b) or argument reasoning comprehension (Niven and Kao, 2019; Branco et al., 2021) with only the claim, choosing a plausible story ending without seeing the story (Cai et al., 2017), question answering using a positional bias (Jia and Liang, 2017; Kaushik and Lipton, 2018), etc.

The spurious correlation problems in information extraction tasks are still an under-explored area. Despite most recent studies on NER (Zhang et al., 2021) and relation extraction (Wang et al., 2022), this work represents the first attempt to diagnose spurious correlations in entity typing, for which we comprehensively analyzed various types of causes for biases and provided a dedicated debiasing method. We also conducted a comprehensive comparison with various alternatives based on resampling (Le Bras et al., 2020), reweighting (Clark et al., 2019; Karimi Mahabadi et al., 2020) and counterfactual inference (Wang et al., 2022).

5 Conclusions

To comprehensively investigate the faithfulness and reliability of entity typing methods, we systematically define six kinds of model biases that are reflected mainly from spurious correlations. In addition to diagnosing the biases on representative models using benchmark data, we also present a counterfactual data augmentation approach that helps improve the generalization of different entity typing models with consistently better performance on both original and debiased test sets.

Limitations

There are two important caveats to this work. First, for instances identified with a particular bias by the PLM, we do not guarantee all typing models would exploit spurious correlations on it. To the best of our knowledge, entity typing models with spurious correlation ablated and mitigated do not yet exist. Although we observe significant performance differences between the original biased instances and the crafted debiased counterparts from existing entity typing models, we hope future work would pay attention to spurious correlations, and develop models with improved robustness and generalization performance. Second, although biases defined in this work comprehensively cover six aspects, but still they may not exhaust all kinds of biased prediction in entity typing. In our study we only tried our best effort to study the most noteworthy and typical biases with which models may inflate performance by leveraging corresponding spurious correlations. At the same time, appeal for more research efforts to complete our understanding with more biases investigated. In addition, the studied model biases are representative to the widely practiced classification-based typing paradigm. There are effects in the most recent NLI-based or bi-encoder-based methods (Li et al., 2022; Huang et al., 2022), which require further analysis.

Ethical Consideration

We acknowledge the importance of ethical considerations in language technologies and would like to point the reader to the following concern. Gender is a spectrum and we respect all gender identities, e.g., nonbinary, genderfluid, polygender, omnigender, etc. To craft instances free from pronoun bias, we substitute the gender pronouns with their referred names in contexts if they exist, or random masculine/feminine given names otherwise. This is due to the lack of entity typing datasets going beyond binarism for pronoun mentions such as they/them/theirs, ze/hir/hir, etc. Nevertheless, we support the rise of alternative neutral pronoun expressions and look forward to the development of non-binary inclusive datasets and technologies. In the meantime, although our techniques do not introduce or exaggerate possible gender bias in the original experimental data, in cases where such biases pre-exist in those data, additional gender neuralization techniques would be needed in order for such biases to be mitigated.

Acknowledgement

We appreciate the anonymous reviewers for their insightful comments and suggestions. This material is partly supported by the National Science Foundation of United States Grant IIS 2105329 and a Cisco Research Award. Nan Xu and Fei Wang are supported by the Annenberg Fellowships. Bangzheng Li is supported by the USC Provost's Ph.D. Fellowship. Mingtao Dong is supported by the USC Provost's Undergrad Research Fellowship.

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending:strong neural baselines for the ROC story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 616–622, Vancouver, Canada. Association for Computational Linguistics.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.

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.

Hongliang Dai, Donghong Du, Xin Li, and Yangqiu Song. 2019. Improving fine-grained entity typing with entity linking. 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 6210–6215, Hong Kong, China. Association for Computational Linguistics.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-fine entity typing with weak supervision from a masked language model. 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 1790–1799, Online. Association for Computational Linguistics.

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021a. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021b. Towards interpreting and mitigating shortcut learning behavior of nlu models. *arXiv preprint arXiv:2103.06922*.

Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. 2014. Context-dependent fine-grained entity type tagging. *arXiv preprint arXiv:1412.1820*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Feng Hou, Ruili Wang, Jun He, and Yi Zhou. 2020. Improving entity linking through semantic reinforced entity embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6843–6848, Online. Association for Computational Linguistics.

James Y. Huang, Bangzheng Li, Jiashu Xu, and Muhao Chen. 2022. Unified semantic typing with meaningful label inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654, Seattle, United States. Association for Computational Linguistics.

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.

Hailong Jin, Lei Hou, Juanzi Li, and Tiansi Dong. 2019. Fine-grained entity typing via hierarchical multi graph convolutional networks. 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 4969–4978, Hong Kong, China. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a

critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Duong Le and Thien Huu Nguyen. 2021. Fine-grained event trigger detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2745–2752, Online. Association for Computational Linguistics.

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.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022. Ultra-fine entity typing with indirect supervision from natural language inference. *Transactions of the Association for Computational Linguistics*, 10:607–622.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Qing Liu, Hongyu Lin, Xinyan Xiao, Xianpei Han, Le Sun, and Hua Wu. 2021. Fine-grained entity typing via label reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4611–4622, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

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.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Yasumasa Onoe, Michael Boratko, Andrew McCallum, and Greg Durrett. 2021. Modeling fine-grained entity types with box embeddings. 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 2051–2064, Online. Association for Computational Linguistics.

Yasumasa Onoe and Greg Durrett. 2019. Learning to denoise distantly-labeled data for entity typing. 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 2407–2417, Minneapolis, Minnesota. Association for Computational Linguistics.

Yasumasa Onoe and Greg Durrett. 2020. Interpretable entity representations through large-scale typing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 612–624, Online. Association for Computational Linguistics.

Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron C Wallace. 2021. Combining feature and instance attribution to detect artifacts. *arXiv preprint arXiv:2107.00323*.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. 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 5434–5445, Online. Association for Computational Linguistics.

Xiang Ren, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, and Jiawei Han. 2016. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1825–1834.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Li Tianran, Yang Erguang, Zhang Yujie, Chen Yufeng, and Xu Jinan. 2021. Improving entity linking by encoding type information into entity embeddings. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1087–1095, Huhhot, China. Chinese Information Processing Society of China.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120,

Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. 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.

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

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. *NAACL*.

Zhao Wang and Aron Culotta. 2020. 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.

Wenhan Xiong, Jiawei Wu, Deren Lei, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Imposing label-relational inductive bias for extremely fine-grained entity typing. 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 773–784, Minneapolis, Minnesota. Association for Computational Linguistics.

Peng Xu and Denilson Barbosa. 2018. Neural fine-grained entity type classification with hierarchy-aware loss. 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 16–25, New Orleans, Louisiana. Association for Computational Linguistics.

Dani Yogatama, Daniel Gillick, and Nevena Lazic. 2015. Embedding methods for fine grained entity type classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 291–296, Beijing, China. Association for Computational Linguistics.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021. De-biasing distantly supervised named entity recognition via causal intervention. 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 4803–4813, Online. Association for Computational Linguistics.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL-IJCNLP)*.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Additional Details about Mitigating Spurious Correlations

Lexical Overlapping Bias We consider the following sentence as an instance: "Deutsche Bank would neither confirm nor deny the ***discharge*** of the two executives, and it also would not specify who was the target of the alleged spying", annotated with types *dismissal, discharge, leave, termination*. Since "discharge" shows up both in the mention and the true labels, we perform word substitutions with synonym candidates from 20 synsets found in WordNet. We show a few synsets with popular senses as follows:

Synset I: (*the termination of someone's employment*) dismissal, dismission, discharge, firing, liberation, release, sack, sacking

SynsetII: (*a substance that is emitted or released*) discharge, emission

SynsetIII: (*a formal written statement of relinquishment*) release, waiver, discharge

Synonyms that share high word similarities with the true labels are removed to avoid creating new lexical overlapping bias features, e.g., *dismissal, discharge* from Synset I, *discharge* from Synset II and Synset III. To guarantee the semantic consistency of the new sentence and the fidelity of true labels to type the new mention, we leverage available word sense disambiguation models to preserve synonyms from the synset that is most consistent with the sense used in the original sentence: *dismission, firing, liberation, release, sack, sacking* from Synset I are finally selected to substitute "discharge". As shown in T4 of Tab. 1, without training on the debiased set, *MLMET* no longer predicts the overlapped type "day", but some surface word "period" instead.

Named Entity Bias Compared with the politician *Benjamin Netanyahu*, the PLM can hardly infer the impression of the singer *Jintara Poonlarp* on the public. Particularly, only general types to describe person named entities are predicted in S5: *person, human, woman*. We then consider *Benjamin Netanyahu* as a biased named entity containing much prior knowledge, while *Jintara Poonlarp* as an unbiased named entity without much type-relevant information revealed. After substituting *Benjamin Netanyahu* with *Jintara Poonlarp* in T6, *MLMET* can hardly infer the political role of the new mention by analyzing its connection with the politician (*Amin al-Husseini*, Palestinian Arab nationalist and Muslim leader in Mandatory Palestine[1]) and political description ("masterminds" and "Holocaust") in the context. *MLMET* even crashes with some out-of-context predictions: *scholar, writer*.

Pronoun Bias As shown in the original instance T7 of Tab. 1, the actual person's name that the pronoun mention "Her" refers to is not provided in the current sentence. As a result, a random feminine name, "Judith" is assumed to be the referred entity and substitutes the pronoun mention as a new sentence in S6. Considering the ridiculously wrong types predicted by *RoBERTa* such as *bird* and *cat*, we include this new instance in the debiased set and expect the entity modeling training on this kind of instances to infer person name types as accurate as pronoun types. Beforehand, we test on the newly crafted instance without counterfactual augmented training, and observe huge performance drop after pronoun concretization: types related to the name's gender attribute such as *woman* and *female* are missing, let alone the types requiring fully context understanding such as *actress*.

Dependency Bias For instance T9 in Tab. 2, we show their mention word dependency analysis in S6 and predictions on the perturbed instance in T10. Without distractions from other dependent words in the new mention, *MLMET* spares no effort to infer types of the target entity "whale" with the correct prediction *subject*. Motivated by the improved performance when the mention headword is specifically provided, we believe entity typing models can actively learn to capture target entity among mention words when both original sentences and their debiased counterparts are given during training. In such augmented training regime, the entity typing model is expected achieve robust performance on new sentences bearing distractions from dependent words in mentions.

A.2 Implementation Details

We adopt the released checkpoints of RoBERTa-large (Liu et al., 2019) as the PLM to identify biased instances. To perform masked fill-in, we adopt the top 10 predictions and filter out non-type words as the predicted types. We recognize potentially

[1] https://en.wikipedia.org/wiki/Amin_al-Husseini

biased samples based on PLM predictions based on the following criteria. *1) mention-context bias*: instances are considered biased if the PLM can predict the type labels with the F1 score above 0.5 when only the mention is provided; *2) named entity bias*: instances are considered biased if the PLM can predict types labels with the F1 score above 0.5 when only the named entity is given; *3) lexical overlapping bias*: instances are considered biased if the PLM makes predictions with the F1 score below 0.5 after substituting overlapped words with their semantically similar words; *4) pronoun bias*: for pronouns without coreferenced entities detected, we substitute them with 5 random real person names as debiased instances. Instances are considered biased if the PLM makes predictions with the F1 score below 0.5 after real name substitution. We mainly use 0.5 as the threshold to distinguish biased samples from unbiased, since the SOTA model achieves the F1 score approximating 0.5 on average of the UFET test samples.

To diagnose entity typing models, for those with released checkpoints (*BiLSTM*, *Box4Types*, *LRN*), we directly evaluate on the original (un)biased and crafted debiased instances. We train *LabelGCN* and *MLMET* by ourselves following hyperparameters and training strategies introduced in their papers.

To evaluate various debiasing approaches, we train entity typing models using checkpoints training on the original dataset as the warm start with the same hyperparameter sets.

We run experiments on a commodity server with a GeForce RTX 2080 GPU. It takes about 4 hours to train one entity typing model on average and 2 minutes for inference on the UFET test set.

A.3 OntoNotes Experiments

We diagnose entity typing models and the effectiveness of the proposed counterfactual augmented approach on OntoNotes (Gillick et al., 2014). The original dataset contains $251,309$ instances automatically annotated by linking identified entity mentions to Freebase profiles for training, and $11,165$ manually annotated instances: $2,202$ for validation and $8,963$ for testing, respectively. Its label space is constituted of 89 types organized into a hierarchy, e.g., /person (level 1), /person/artist (level 2), /person/artist/actor (level 3). We adopt the set augmented by (Choi et al., 2018) for model training: $793,487$ instances with distant supervision from Wikipedia definition sentences and head word supervision.

In Tab. 8, we report performance of two representative entity typing models on original biased samples where they are likely to exploit spurious correlations, the perturbed counterparts, as well as performance on unbiased samples. We have the following observations: 1) entity typing models can achieve satisfactory performance when only the mention is provided without context; 2) considering lexical overlapping bias, performance on both biased and unbiased samples identified by the PLM drops a lot after substituting overlapped mention words with their sematically similar words; 3) the performance variation after named entity substitution is evident; 4) models can obtain much better performance on some instances when the headwords are explicitly given without distractions from other words in mentions; 5) performance on instances purely annotated by coarse and fine labels is good in general with around 15% difference in F1 score. Similarly to UFET, models training on OntoNotes may achieve good performance without reasoning on the context, rely on lexical overlapping between mention words and types to make precise predictions, and obtain below-average results on some instances for lack of syntactic structure understanding.

To mitigate spurious correlations, we evaluate the proposed counterfactual augmented approach in Tab. 9. With additional debiased instances for model training, both *BiLSTM* and *MLMET* maintain good performance on the original OntoNotes test set and much higher accuracy on the corresponding debiased test set, leading to improved generalization.

Test Set	BiLSTM			MLMET			LabelGCN			Box4Types			LRN		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
						Mention-Context Bias (↑)									
Biased	.602	.283	.385	.668	.640	.654	.689	.418	.521	.679	.516	586	.666	.425	.519
-Perturb.	.606 (0.7%)	.298 (5.3%)	.400 (3.8%)	.682 (2.1%)	.628 (-1.9%)	.654 (0.0%)	.699 (1.3%)	.398 (-4.9%)	.507 (-2.6%)	.644 (-5.2%)	.486 (-5.8%)	.554 (-5.5%)	.653 (-2.0%)	.434 (2.2%)	.522 (0.5%)
Unbiased	.452	.188	.265	.486	.395	.436	.484	.235	.317	.470	.319	.380	.563	.272	.367
-Perturb.	.450 (-0.5%)	.176 (-6.4%)	.253 (-4.7%)	.453 (-6.8%)	.350 (-11.4%)	.395 (-9.4%)	.459 (-5.2%)	.221 (-6.1%)	.298 (-5.8%)	.446 (-5.1%)	.279 (-12.6%)	.343 (-9.8%)	.509 (-9.6%)	.232 (-14.7%)	.319 (-13.1%)
						Lexical Overlapping Bias (↓)									
Biased	.415	.661	.510	.641	.484	.551	.651	.292	.403	.118	.442	.186	.089	.259	.133
-Perturb.	.040 (-90.4%)	.065 (-90.1%)	.050 (-90.3%)	.411 (-35.8%)	.271 (-44.0%)	.327 (-40.8%)	.338 (-48.1%)	.106 (-63.6%)	.162 (-59.9%)	.074 (-36.9%)	.251 (-43.3%)	.115 (-38.3%)	.066 (-25.8%)	.158 (-39.0%)	.093 (-29.7%)
Unbiased	.278	.667	.392	.522	.567	.544	.667	.300	.414	.167	.667	.267	.067	.333	.111
-Perturb.	.111 (-60.0%)	.333 (-50.0%)	.167 (-57.5%)	.381 (-27.1%)	.533 (-5.9%)	.444 (-18.2%)	.333 (-50.0%)	.067 (-77.8%)	.111 (-73.1%)	.222 (33.3%)	.667 (0.0%)	.333 (25.0%)	.056 (-16.7%)	.333 (0.0%)	.095 (-14.3%)
						Named Entity Bias (↓)									
Biased	.744	.380	.504	.719	.752	.735	.730	.524	.610	.686	.658	.671	.754	.557	.641
-Perturb.	.538 (-27.7%)	.240 (-36.9%)	.332 (-34.0%)	.615 (-14.5%)	.586 (-22.0%)	.600 (-18.4%)	.568 (-22.2%)	.362 (-30.9%)	.442 (-27.5%)	.541 (-21.2%)	.479 (-27.2%)	.508 (-24.4%)	.634 (-15.9%)	.448 (-19.5%)	.525 (-18.0%)
Unbiased	.522	.226	.316	.536	.477	.505	.542	.288	.376	.520	.407	.457	.595	.365	.453
-Perturb.	.613 (17.5%)	.267 (17.7%)	.372 (17.6%)	.582 (8.6%)	.502 (5.4%)	.539 (6.9%)	.651 (20.2%)	.337 (17.3%)	.444 (18.2%)	.538 (3.4%)	.472 (16.1%)	.503 (10.2%)	.629 (5.8%)	.433 (18.6%)	.513 (13.4%)
						Pronoun Bias (↓)									
Biased	.567	.438	.494	.555	.566	.561	.555	.474	.511	.576	.555	.565	.716	.474	.570
-Perturb.	.148 (-74.0%)	.227 (-48.2%)	.179 (-63.8%)	.619 (11.5%)	.455 (-19.6%)	.525 (-6.4%)	.578 (4.2%)	.300 (-36.7%)	.395 (-22.8%)	.481 (-16.5%)	.397 (-28.4%)	.435 (-23.0%)	.776 (8.2%)	.415 (-12.3%)	.541 (-5.2%)
Unbiased	.405	.334	.366	.738	.635	.683	.660	.510	.575	.715	.759	.736	.633	.346	.448
-Perturb.	.106 (-73.9%)	.167 (-50.0%)	.130 (-64.6%)	.735 (-0.5%)	.597 (-6.0%)	.659 (-3.5%)	.703 (6.6%)	.258 (-49.4%)	.378 (-34.4%)	.670 (-6.3%)	.548 (-27.8%)	.603 (-18.1%)	.555 (-12.4%)	.339 (-2.1%)	.421 (-6.0%)
						Dependency Bias									
UFET	.350	.097	.152	.450	.402	.424	.452	.248	.321	.462	.351	.399	.491	.300	.372
-Perturb.	.407 (16.2%)	.386 (297.3%)	.396 (160.5%)	.617 (37.1%)	.520 (29.4%)	.564 (32.9%)	.757 (67.3%)	.392 (58.1%)	.517 (61.2%)	.710 (53.7%)	.482 (37.2%)	.574 (43.9%)	.715 (45.7%)	.445 (48.5%)	.549 (47.4%)
						Overgeneralization Bias									
Coarse	.362	.656	.466	.297	.758	.427	.360	.656	.465	.339	.688	.454	.466	.683	.554
Ultra-fine	.157	.094	.118	.221	.244	.232	.207	.120	.152	.147	.120	.132	.161	.091	.116

Table 7: Performance of all entity typing models evaluated by complete metrics (Prec. for precision, Rec. for recall and F1 for F1 score) on UFET testing samples with(out) distinct bias and their perturbations.

Test set	Mention-Context 1830/7133 (↑)		Lexical Overlapping 234/319 (↓)		Named Entity 1132/2457 (↓)		Dependency 842,544/7129		Overgeneralization 5549/3414	
	BiLSTM	MLMET	BiLSTM	MLMET	BiLSTM	MLMET	BiLSTM	MLMET	BiLSTM	MLMET
Biased	.719	.821	.922	.940	.774	.844	.291	.416	.844	.909
-Perturb.	.698 (-3.0%)	.803 (-2.2%)	.345 (-62.6%)	.473 (-49.7%)	.668 (-13.7%)	.765 (-9.4%)	.808 (178.0%)	.909 (118.5%)	.646	.757
Debiased	.787	.864	.982	.983	.761	.855	-	-	-	-
-Perturb.	.787 (.1%)	.847 (-2.0%)	.407 (-58.6%)	.467 (-52.5%)	.665 (-12.6%)	.759 (-11.3%)	-	-	-	-

Table 8: F1 score of two representative entity typing models on OntoNotes testing samples with(out) distinct biases and their perturbations.

Approach	BiLSTM						MLMET					
	U-Prec.	U-Rec.	U-F1	A-Prec.	A-Rec.	A-F1	U-Prec.	U-Rec.	U-F1	A-Prec.	A-Rec.	A-F1
No Debiasing	**.803**	.744	**.773**	.708	.609	.655	**.890**	.822	**.855**	.805	.706	.753
Augmentation (ours)	.782	***.752***	.767	***.781***	***.711***	***.745***	.777	***.832***	.803	***.828***	***.846***	***.837***

Table 9: Effectiveness of the proposed counterfactual augmented approach on two representative entity typing models when testing on OntoNotes test set (U-) and our counterfactual augmented test set (A-).

Approach	U-Prec.	U-Rec.	U-F1	A-Prec.	A-Rec.	A-F1
		LabelGCN				
No Debiasing	.498	.283	.361	.503	.247	.332
AFLite	***.536***	.238	.329	***.529***	.202	.292
POE	.407	*.327*	*.363*	.379	*.301*	*.335*
Focal	.185	***.439***	.260	.193	***.392***	.259
Learned-mixin	.420	*.316*	.361	.353	*.311*	.331
Learned-mixin+H	.225	*.398*	.287	.212	*.339*	.261
Contrastive (CE)	.483	*.286*	.359	.479	*.272*	*.347*
Contrastive (Cosine)	.453	*.285*	.350	*.516*	*.269*	*.353*
Counterfact. Inf.	.467	*.309*	***.372***	.403	*.291*	*.338*
Augmentation (ours)	.484	*.289*	*.362*	*.524*	*.274*	***.360***
		Box4Types				
No Debiasing	.528	.388	.448	.469	.358	.406
AFLite	***.531***	*.400*	*.456*	*.473*	*.360*	*.409*
POE	.410	***.468***	.437	.347	*.433*	.385
Focal	.407	*.467*	.435	.347	*.435*	.386
Learned-mixin	.508	*.415*	*.457*	.448	*.393*	*.419*
Learned-mixin+H	.463	*.440*	*.451*	.403	*.406*	.404
Contrastive (CE)	.443	*.459*	*.451*	.371	*.563*	*.447*
Contrastive (Cosine)	.472	*.444*	*.458*	.437	***.499***	*.466*
Counterfact. Inf.	*.529*	*.394*	*.452*	.422	*.382*	.401
Augmentation (ours)	.521	*.410*	***.459***	***.504***	*.484*	***.494***
		LRN				
No Debiasing	**.611**	**.334**	**.432**	**.703**	.343	.461
Augmentation (ours)	.553	.328	.412	.619	***.389***	***.478***

Table 10: Effectiveness of different debiasing approaches on remaining entity typing models when testing on UFET test set (U-) and our counterfactual augmented test set (A-). Note that *LRN* predicts types in an autoregressive generative way, it does not provide a fixed logit for each label, hence we can not apply logit-based debiasing approaches to help *LRN* mitigate spurious correlations.