How Fragile is Relation Extraction under Entity Replacements?

Yiwei Wang[†] **Bryan Hooi**[‡] **Fei Wang**[§] **Yujun Cai**[¶] **Yuxuan Liang**[‖]
Wenxuan Zhou[§] **Jing Tang**[‖] **Manjuan Duan**[††] **Muhao Chen**[‡‡]
[†] University of California, Los Angeles [‡] National University of Singapore
[§] University of Southern California [¶] Meta
[‖] Hong Kong University of Science and Technology (Guangzhou)
[††] Amazon [‡‡] University of California, Davis
wangyw_seu@foxmail.com

Abstract

Relation extraction (RE) aims to extract the relations between **entity names** from the **textual context**. In principle, textual context determines the ground-truth relation and the RE models should be able to correctly identify the relations reflected by the textual context. However, existing work has found that the RE models memorize the entity name patterns to make RE predictions while ignoring the textual context. This motivates us to raise the question: "are RE models robust to the entity replacements?" In this work, we operate the random and type-constrained entity replacements over the RE instances in TACRED and evaluate the state-of-the-art RE models under the entity replacements. We observe the 30% - 50% F1 score drops on the state-of-the-art RE models under entity replacements. These results suggest that we need more efforts to develop effective RE models robust to entity replacements. We release the source code at https://github.com/wangywUST/RobustRE.

1 Introduction

Recent literature has shown that the sentence-level relation extraction (RE) models may overly rely on entity names for RE instead of reasoning from the textual context (Peng et al., 2020; Wang et al., 2022). This problem is also known as *entity bias*: the spurious correlation between entity names and relations (Longpre et al., 2021; Qian et al., 2021; Xu et al., 2022; Wang et al., 2022). This motivates us to raise a question: "how robust are RE models under entity replacements?"

Entity bias degrades the RE models' generalization, such that the entity names can mislead the models to make wrong predictions. However, a seemingly conflicting phenomenon is the high (in-distribution) accuracy of RE models on the standard benchmarks, such as TACRED. In our work, we find that these benchmarks are prone to have shortcuts from entity names to ground-truth relations (see Fig. 2), low entity diversity, and a large portion of incorrect entity annotations. These issues suggest that, given the presence of entity bias, the current benchmarks are not challenging enough to evaluate the generalization of RE in practice.

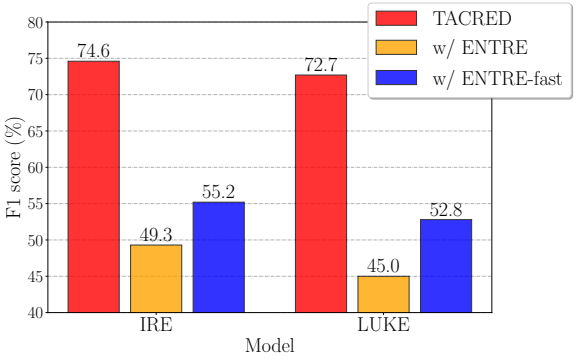

Figure 1: The performance of state-of-the-art RE models drop a lot under entity replacements (ENTRE).

Evaluating RE with valid instances of more comprehensive entities is non-trivial. It requires us to collect many sentences containing comprehensive entities and carefully label the relations. Both the text collection and annotations are time-consuming and expensive. Instead, in our work, we aim to efficiently produce rich valid RE instances with comprehensive entities based on the carefully designed entity replacements. Most existing methods for evaluating the generalizability of NLP focus on sentence classification (Jin et al., 2020; Li et al., 2020; Minervini and Riedel, 2018) and question answering (Jia and Liang, 2017; Ribeiro et al., 2018; Gan and Ng, 2019), but these methods lack special designs to seize on the entity bias in RE.

In this work, we propose a **type-constrained** and **random** entity replacement method: ENTRE. **Type-constrained** means we replace the named entity in the type `[PERSON]` or `[ORGANIZATION]` with the new entity belonging to the same type as the original entity. **Random** means we randomly select the entity names from a Wikipedia entity

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 414–423
December 6–7, 2023.

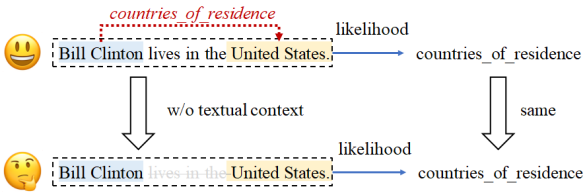

Figure 2: TACRED offers many shortcuts from entity names to ground-truth relations in the test set, where the model predicts the correct relation even when only given the entity names, despite all textual context being removed. As a result, TACRED is not challenging enough to measure the generalization under entity bias.

lexicon that consists of 24,933 organizations and 902,007 person entities for replacements. These two principles guarantee the effectiveness of entity replacement to produce valid and diverse RE instances.

We apply ENTRE to TACRED and evaluate the RE models on the instances with replaced entity names. We analyze the RE models under entity replacements in order to answer four research questions: (Q1) How do the strong RE models perform under entity replacements? (Q2) Does ENTRE reduce prediction shortcuts from entity names to the ground-truth relations? (Q3) Does ENTRE improve the entity diversity? (Q4) How to improve the robustness of RE?

We observe several key findings. First, the strong RE models LUKE (Yamada et al., 2020) and IRE (Zhou and Chen, 2021) tend to memorize entity-relation patterns to infer the relation instead of reasoning based on the textual context that actually describes the relation. This phenomenon causes the model to be brittle to entity replacements, resulting in a significant performance drop of 30% - 50% in terms of the F1 score. Second, ENTRE reduces the shortcuts by more than 50% on many relations, and improves the subject name diversity by more than 25 times compared to TACRED. Third, the recent causal inference approach CoRE (Wang et al., 2022) improves the robustness at a higher magnitude than other methods.

For the easy use of ENTRE, we provide a challenging RE benchmark built by ENTRE: ENTRED, which consists of the TACRED test set instances with the entity names replaced by ENTRE. We believe the proposed ENTRE and benchmark ENTRED will benefit future research toward improving the RE robustness.

Index: 7523
Text: He figured that he would sell his home before the interest rate on the loan, taken out from Countrywide Financial (subject, organization), now owned by Bank of America (object, person), reset at a higher level.

Index: 4190
Text: It traveled to Rice University, where de Menil and his (subject, organization) wife, Dominique de Menil, who later founded the Menil Collection (object, organization), ran the art museum.

Figure 3: Two examples of incorrect entity annotations in TACRED.

2 Analysis of Entity Names in TACRED

Before introducing ENTRE, we first analyze the existing popular RE datasets. Our analysis is focused on the following three perspectives: 1) the correctness of entity name annotations; 2) the diversity of entity names; 3) the prediction shortcuts from entity names to the ground-truth relations.

In the popular TACRED (Zhang et al., 2017), TACREV (Alt et al., 2020), and Re-TACRED (Stoica et al., 2021) datasets, we find that: first, there exist some portion of incorrect entity name annotations; second, many entity names are reused more than one hundred times across instances; third, the entity names in more than 70% of the instances act as shortcuts to the ground-truth relations. We introduce the details as follows.

2.1 Incorrect Entity Annotations

In the TACRED (Zhang et al., 2017), TACREV (Alt et al., 2020), and Re-TACRED (Stoica et al., 2021) datasets, there exist quite a few incorrect entity annotations. To detect these incorrect entity annotations, we use a BERT based NER model (Devlin et al., 2019) to automatically annotate the subject and object entity names in the TACRED dataset. Then, we conduct manual investigation on the entities where the NER annotations are different the original TACRED annotations. We find that more than 10% of the test instances contain incorrect entity annotations.[1] We present two examples in Fig. 3. Using these mistaken entity annotations to evaluate the RE models compromises our goal of correctly measuring RE performance.

[1]Including both incorrect span and type annotations.

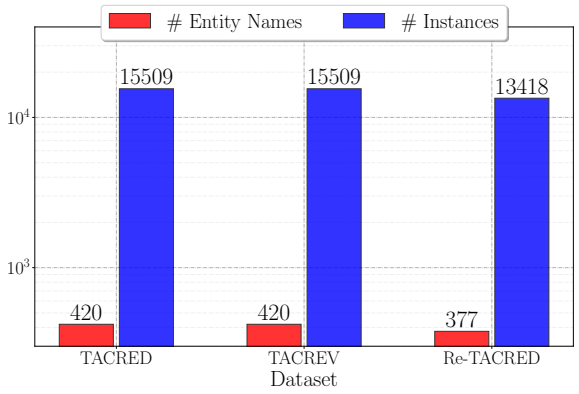

Figure 4: The number of different subject entity names (red) is much lower than the number of instances (blue) in the test sets of the TACRED, TACREV, and Re-TACRED datasets. In other words, the diversity of entity names in these datasets' test sets is limited.

2.2 Diversity of Entity Names

The TACRED, TACREV, and Re-TACRED datasets have a low diversity of entity names: most entity names repeatedly appear in a large portion of instances (see Fig. 4). In the TACRED datasets, there are only 420 entity names repeatedly appearing as 15509 instances' subjects. For example, *"ShopperTrak"*, as the subject, has repeatedly appeared as the subject entity in 270 instances. This heavily repeated use of entity names increases the risk that RE relies on entity bias to make RE predictions. Also, with these benchmarks, it is impossible to comprehensively evaluate the generalization of RE models on a diverse set of entity names to imitate real-world scenarios.

2.3 Causal Inference for Entity Bias

We follow the prior work (Wang et al., 2022) to analyze the entity bias based on causal inference. (Wang et al., 2022) builds the causal graph of RE as a directed acyclic graph: $(E, X) \rightarrow Y$ in Figure 5. X is the input text, E denotes the entity mentions, and Y is the relation extraction result. On the edges $(X, E) \rightarrow Y$, the RE model encodes E and X to predict the relation Y.

Based on the causal graph displayed in Figure 5, we can diagnose whether the entities have shortcuts to relation. Wang et al. (2022) distill the entity bias by counterfactual analysis, which assigns the hypothetical combination of values to variables in a way that is counter to the empirical evidence obtained from data. We mask the tokens in X to conduct the intervention $X = \bar{x}$ on X, while keeping the variable E as the original entity mentions e. In

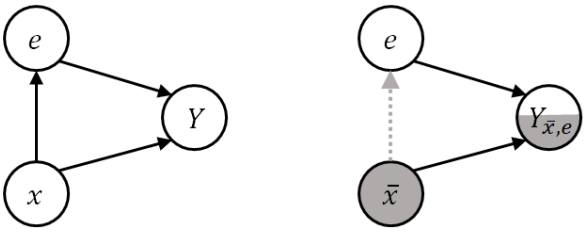

Figure 5: The original causal graph of RE models (left) together with its counterfactual alternatives for the entity bias (right). The shading indicates the mask of corresponding variables.

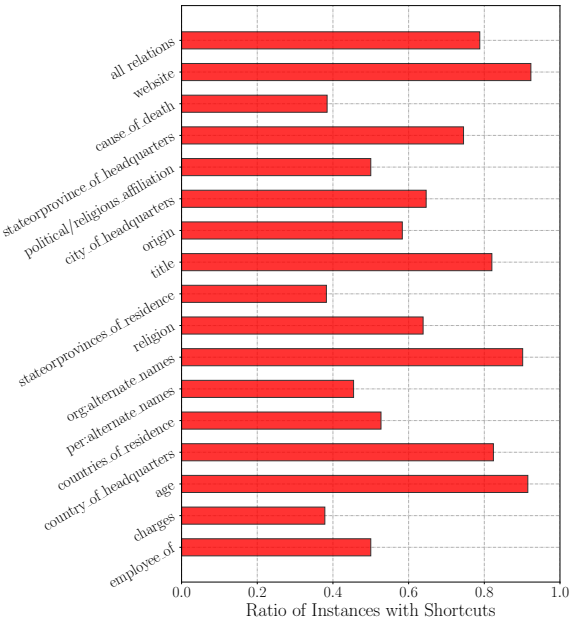

Figure 6: The ratio of instances with shortcuts (the entity bias is as same as the ground truth relation) in the TACRED test set.

this way, the textual context is removed and the entity information is maintained. Accordingly, the counterfactual prediction is denoted as $Y_{\bar{x},e}$ (see Figure 5). $Y_{\bar{x},e}$ refers to the output, i.e., a probability distribution or a logit vector, where only the entity mentions are given.

2.4 Shortcuts to the Ground-Truth Relations

Existing work has found that the popular RE benchmarks' test sets provide abundant shortcuts from entity names to ground-truth relations (Wang et al., 2022; Peng et al., 2020). In other words, on many instances, the model need not "extract" the relation from the textual context but can infer the correct prediction directly through shortcuts from entities.

To verify these observations, we conduct a preliminary study of the shortcuts using the strong RE model LUKE (Yamada et al., 2020) on the TA-

CRED dataset. We first compute the instance-wise relation extraction result in the TACRED's test set. Then, we analyze the shortcuts from entity names to the relations based on causal inference (see details in Sec. 2.3). We find that there exists a large portion of instances having shortcuts from entity names to the ground-truth relations. We visualize the ratio of instances that present shortcuts in different relations in Fig. 6. Last but not least, we observe similar phenomena on other models and TACREV, Re-TACRED datasets as well.

The analyses suggest that these benchmarks do not accurately evaluate the "extraction" capability of RE models without the shortcuts from entity names. In other words, these popular benchmarks are not challenging enough to evaluate whether the RE models can extract the correct relations from the textual context. In our work, we replace the entity names to reduce the shortcuts, to mitigate the possibility that RE models rely on the shortcut of entity bias to achieve over-optimistically high RE performance. Our ENTRE is able to better simulate real-world scenarios with fewer shortcuts and higher entity diversity, which provides a better evaluation of the generalization of RE models.

3 Entity Replacement for RE

We present ENTRE: a simple yet effective procedure to generate high-quality RE instances with entity replacements. ENTRE replaces entity names in the RE instances in a random and type-constrained way. We apply ENTRE to the test set of TACRED to evaluate the state-of-the-art RE models' robustness under entity replacements.

3.1 Targetting the Entities for Replacements

We desire entity replacements to not affect the soundness of language. As we have analyzed in Sec. 2.1, there exists a significant amount of incorrect entity annotations in TACRED. To handle these incorrect entity annotations, we use a BERT based NER model (Devlin et al., 2019) to re-annotate the entities in the TACRED test set. Then, we further conduct a manual investigation over the entity annotations. We filter out incorrectly annotated instances and only replace the named entities. This prevents our entity name replacements from altering the ground-truth relation labels.

Besides the incorrect entity annotations, there are also some entities for which replacement may inevitably cause noise. For example, some entities belong to the `[MISC]` (miscellaneous) class. If we replace a `[MISC]` entity with another `[MISC]` one, it is likely that we will break the semantics of the original sentence. In contrast, replacing the `[PERSON]` and `[ORGANIZATION]` entities with those belonging to the same type generally do not affect the ground-truth relations. We notice that all the instances in TACRED have a `[PERSON]` or `[ORGANIZATION]` entity as the subject or object. Therefore, in our work, we focus on replacing the `[PERSON]` and `[ORGANIZATION]` entities.

3.2 Large Lexicon of Entities

We propose the following standards for selecting the new entity names for replacements:

1. The new entity belongs to the same type as the replaced one.
2. The new entity exists in the real world.
3. The new entity names are more diverse.

These three standards contribute to making the resulting instances *natural* – i.e., containing real, valid entities that are of the same class as the original entities, and are linguistically sound; *challenging* – i.e., the new entities may not offer shortcuts to the model, which cannot easily get the correct extraction result by seeing only the entity names and *comprehensive* – i.e., the robustness of RE is evaluated on a more diverse set of entities.

To satisfy the above standards, we first build up a large entity name lexicon to provide the new entity names for replacements. The size of the entity lexicon determines the diversity of entity names in our new RE benchmark ENTRED. Also, a larger entity name lexicon can help us to evaluate the generalization of RE models on more out-of-domain entity names in test time. Therefore, in addition to the entity names appearing in the TACRED, we collect the entity names from Wikipedia belonging to the category of person and organization to enrich the entity name corpus. Overall, we collect 24,933 organization and 902,007 person names from Wikipedia[2] to build a large entity lexicon.

3.3 Entity Replacements

Based on the constructed entity lexicon, we propose ENTRE: a type-constrained and random entity replacement method. **Type-constrained** means we

[2] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

Benchmark	TACRED	ENTRED
# Sentences	15,509	12,419
# Tokens	539,306	457,121

Table 1: Statistics of the TACRED and ENTRED benchmarks.

replace the named entity in the type [PERSON] or [ORGANIZATION] with the new entity belonging to the same type as the original entity. **Random** means we randomly select the entity names from our entity lexicon that consists of 24,933 organizations and 902,007 person entities for replacements. These two principles guarantee the effectiveness of entity replacement to produce valid RE instances. We iterate over TACRED instances and replace the entity names. We summarize ENTRE as the following pipeline:

1. Collecting the instances with predictions as same as the ground-truth relation.

2. Replace the entity names for the collected entities in Step 1. Return to step 1.

The above steps can be repeated for many times, and a higher repetition time leads to a higher level of the adversary. We can stop the repeating until all the entities in the lexicon have been used. But that will induce too long running time. Therefore, in our work, we set the maximum number of repetitions as 200 by default.

Step 1 requires the inference on many test instances, which is time-consuming. Considering that the F1 score's calculation of RE takes the "no_relation" as the background class, we can alternatively collect the instances not belonging to the "no_relation" class in Step 1. We denote such an alternate as ENTRE-fast, which saves 90% evaluation time in the experiments.

Both the ENTRE and ENTRE-fast are dataset-agnostic and model-agnostics. In other words, we can apply ENTRE and ENTRE-fast to many RE datasets to evaluate any RE model. In this work, to enable the easy use of ENTRE, we create the challenging RE benchmark ENTRED by applying ENTRE on the test set of TACRED. The overall statistics of ENTRED are shown in Table 3, alongside the statistics of the original TACRED dataset. The number of sentences in ENTRED is slightly smaller than that in TACRED because we filter out the incorrectly annotated instances. We showcase ENTRE using TACRED in this paper because of its popularity on evaluating RE models and comprehensive relation-type coverage. However, our ENTRE can be applied to other RE datasets.

4 Experiments

In this section, we investigate ENTRE and use it to evaluate the robustness of the strong RE models LUKE (Yamada et al., 2020), IRE (Zhou and Chen, 2021), and other methods that can improve the robustness of RE. Our experimental settings closely follow those of previous work (Zhang et al., 2017; Zhou and Chen, 2021; Nan et al., 2021) to ensure a fair comparison. We organize our results and analysis as four main research questions and their answers.

Q1: How robust is relation extraction?

Main Results We evaluate the robustness of the state-of-the-art RE models LUKE (Yamada et al., 2020) and IRE (Zhou and Chen, 2021) under entity replacements. Our experimental settings closely follow those of previous work (Zhang et al., 2017; Zhou and Chen, 2021; Nan et al., 2021) to ensure a fair comparison. We visualize the empirical results in Fig. 1. We observe that the 30% - 50% drops in terms of F1 scores happen on the state-of-the-art RE models after entity replacements. These results suggest that there remains a large gap between the current research and the really effective RE models robust to entity replacements.

We compare the F1 scores on TACRED and ENTRED, the challenging RE benchmark produced by our ENTRE, in Table 2. We can see that the state-of-the-art LUKE has a significant performance drop in our challenging ENTRED; there is a 44% relative decrease (in the models' F1) in ENTRED as compared to their results before entity replacements.

Case Study We conduct case studies to empirically examine the effects of our entity replacements of ENTRE. Table 3 gives a qualitative comparison example between the RE results on TACRED and our ENTRED. The results show that our ENTRE misleads the strong RE model LUKE to predict incorrect relations. For example, given the TACRED instance "*Finance Ministry spokesperson Chileshe Kandeta who confirmed this on Sunday said Magande signed a loan agreement of 31 million dollars with the ADF for the country 's Poverty Reduction Budget Support.*", there is no relation between the

Method	TACRED	TACRED w/ ENTRE (Ours)	Δ
LUKE (Yamada et al., 2020)	72.7	45.0	↓ 44%
w/ Resample (Burnaev et al., 2015)	73.1	45.8	↓ 37%
w/ Entity Mask (w/o name, w/o type) (Zhang et al., 2017)	21.3	21.0	↓ 1%
w/ Entity Mask (w/o name, w/ type) (Zhang et al., 2017)	44.9	45.9	↑ 2%
w/ Entity Mask (w/ name, w/ type) (Zhang et al., 2017)	72.3	61.2	↓ 15%
w/ Focal (Lin et al., 2017)	72.9	47.1	↓ 35%
w/ CoRE (Wang et al., 2022)	**74.6**	**61.7**	↓ 17%
IRE (Zhou and Chen, 2021)	74.6	49.3	↓ 34%
w/ Resample (Burnaev et al., 2015)	73.9	49.6	↓ 33%
w/ Entity Mask (w/o name, w/o type) (Zhang et al., 2017)	22.0	21.8	↓ 1%
w/ Entity Mask (w/o name, w/ type) (Zhang et al., 2017)	60.9	61.3	↑ 1%
w/ Entity Mask (w/ name, w/ type) (Zhang et al., 2017)	74.6	49.3	↓ 34%
w/ Focal (Lin et al., 2017)	74.1	49.5	↓ 32%
w/ CoRE (Wang et al., 2022)	**74.7**	**64.2**	↓ 14%

Table 2: F1 scores (%) and the performance dropping of RE on the test sets of TACRED and our ENTRED. The best results in each column are highlighted in **bold** font. We additionally report the performance drop (%) compared with the performance on the original TACRED dataset.

subject and object existing in the text. After the entity replacement, LUKE believes that the relation between them is "*members*".

The entity bias can account for this result, where given only the entity mentions *American Association of University Women* and *Willingboro Chapter*, the RE model returns the relation "*members*" without any textual context. This implies that the model makes the prediction for the original input relying on the entity mentions, which leads to the wrong RE prediction. In our work, we replace the original entities with the new ones that convey the entity bias different from the ground-truth label to test the generalization of RE models under entity bias.

Memorizing or Reasoning? We propose ENTRE to test the ability to use the textual context to infer the relations. As the entity replacements of ENTRE do not affect the ground-truth relations, RE models should be robust against entity name changes. However, we observe the large performance drops from our entity replacements.

Therefore, we conclude that the strong RE model LUKE is apt to memorize the entity name patterns for predicting relations and is more brittle when the entities that convey the biases are different from the ground-truth relations existing in the input text. To make RE models more robust, we believe an important future direction is to develop context-based reasoning approaches, taking advantage of inductive biases on the textual context that determines the relations.

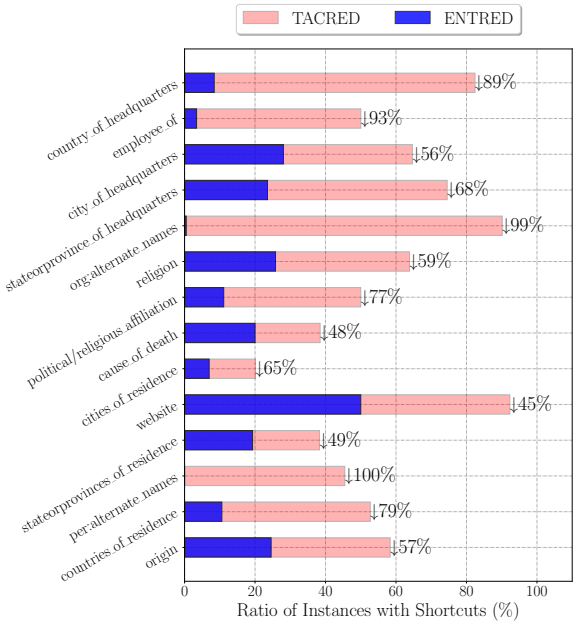

Figure 7: ENTRE significantly reduces the ratio of instances with shortcuts (the entity bias is as same as the ground truth relation) compared with TACRED.

? *Q2: Does* ENTRE *reduce shortcuts?*

ENTRE leads to fewer shortcuts from entity names to ground-truth relations We perform causal inference over ENTRED to analyze how many instances have shortcuts from entity names to the ground-truth relations after the entity replacements. We present the comparison of the shortcut

Original Instance	Original Prediction	New Entity Names	New Prediction
Finance Ministry spokesperson Chileshe Kandeta who confirmed this on Sunday said Magande signed a loan agreement of 31 million dollars with the ADF for the country's Poverty Reduction Budget Support.	no_relation ✓	American Association of University Women, Willingboro Chapter	members ✗
John Graham, a 55-year-old man from Canada, is accused of shooting Aquash in the head and leaving her to die on the Pine Ridge reservation in South Dakota.	stateorprovince_of_death ✓	Liu Shaozhuo, South Dakota	no_relation ✗
After the staffing firm Hollister Inc lost 20 of its 85 employees, it gave up nearly a third of its 3,750-square-foot Burlington office, allowing the property owner to put up a dividing wall to create a space for another tenant.	number_of_employees/members ✓	Yoruba Academy, 85	alternate_names ✗
Kercher 's mother, Arline Kercher, tells court in emotional testimony that she will never get over her daughter 's brutal death.	children ✓	Sanju Yadav, Matti Koistinen	no_relation✗
Lt. Assaf Ramon, the son of Israel's first astronaut, Col. Ilan Ramon, who died in the space shuttle Columbia disaster in 2003, was killed Sunday when an F16-A plane he was piloting crashed in the hills south of Hebron in the West Bank.	children ✓	Aaron Morgan, Ángel Guillermo Heredia Hernández	no_relation ✗
Police have released scant information about the killing of 61-year-old Carol Daniels, whose body was found Sunday inside the Christ Holy Sanctified Church, a weather-beaten building on a rundown block near downtown Anadarko in southwest Oklahoma.	stateorprovince_of_death ✓	Mao Weiming, Oklahoma	no_relation ✗

Table 3: A case study for LUKE on the relation extraction benchmark TACRED and our ENTRED. Underlines and wavy lines highlight the subject and object entities respectively. We report the original prediction, the new entity names for replacements and the prediction in ENTRED.

ratio on ENTRED and TACRED on different relations in Fig. 7. We observe that ENTRED greatly reduces the shortcuts for more than 50% instances on most relations. As a result, when being evaluated using ENTRED, RE models have to extract the informative signals describing the ground-truth relations from the textual context, rather than rely on the shortcuts from the entity names.

? *Q3: Does* ENTRE *improve diversity?*

Comparison between ENTRED and existing benchmarks. As we have analyzed in Sec. 2.1, the diversity of entity names in the existing benchmarks TACRED, TACREV and Re-TACRED are rather limited. These limitations hinder the evaluation of the generalization and generalization of RE. In our work, thanks to our larger lexicon built from the Wikipedia entity names, our ENTRED have much higher diversity than the TACRED and Re-TACRED, as shown in Fig. 8. With these diverse entity names, ENTRED is able to evaluate the performance of RE models on a larger scale of diverse entities, which better imitates the real–world scenario.

? *Q4: How to improve the generalization?*

Methods In our work, we consider the following methods to improve the generalization of RE: (1) **Focal** (Lin et al., 2017) adaptively reweights the losses of different instances so as to focus on the hard ones. (2) **Resample** (Burnaev et al., 2015) up-samples rare categories by the inversed sample

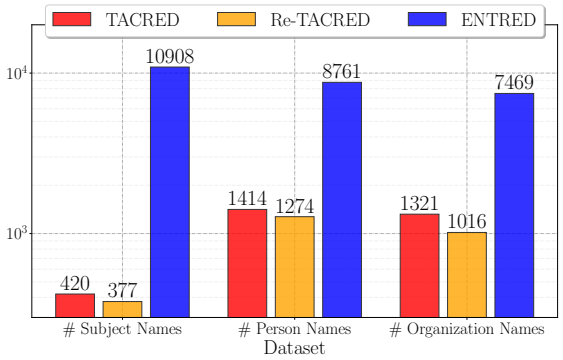

Figure 8: The number of subject entity names, person entity names, and organization entity names in the test set of TACRED (red) and ENTRED (blue).

fraction during training. (3) **Entity Mask** (Zhang et al., 2017): masks the entity mentions with special tokens to reduce the over-fitting on entities. (4) **CoRE** (Wang et al., 2022) is a causal inference based method that mitigates entity bias.

Results & Analysis The results of the above methods on the RE model are shown in Table 2. The recently proposed causal inference based debiasing method CoRE offers the best improvements against our entity replacements ((45.0% → 61.7%)). We conjecture that this is because it mitigates the biasing signals from entity names, which enhances its entity-level generalization ability and makes RE models focus more on the textual context for inference, resulting in a better generalization under entity name replacements. Other methods, however, lead to lower improvements for LUKE, potentially because they cannot effectively capture the biased patterns between relations and entity names.

5 Related Work

Relation extraction (RE) is a sub-task of information extraction that aims to identify semantic relations between entities from natural language text (Zhang et al., 2017). RE is the key component for building relation knowledge graphs, and it is of crucial significance to natural language processing applications such as structured search, sentiment analysis, question answering, and summarization (Huang and Wang, 2017). Early research efforts (Nguyen and Grishman, 2015; Wang et al., 2016; Zhang et al., 2017) train RE models from scratch based on lexicon-level features. The recent RE work fine-tunes pretrained language models (PLMs; Devlin et al. 2019; Liu et al. 2019). For example, K-Adapter (Wang et al., 2020) fixes the parameters of the PLM and uses feature adapters to infuse factual and linguistic knowledge. Recent work focuses on utilizing the entity information for RE (Zhou and Chen, 2021; Yamada et al., 2020), but this leaks superficial and spurious clues about the relations (Zhang et al., 2018). Despite the biases in existing RE models, scarce work has discussed the spurious correlation between entity mentions and relations that cause such biases. Our work builds an automated pipeline to generate natural instances with fewer shortcuts and larger coverage at scale to reflect the serious effects of entity bias on the RE models.

There is also work in other domains aiming to evaluate models' generalization to perturbed inputs. For example, Jia and Liang (2017) attacks reading comprehension models by adding word sequences to the input. Gan and Ng (2019) and Iyyer et al. (2018) paraphrase the input to test models' oversensitivity. Jones et al. (2020) target adversarial typos. Si et al. (2021) propose a benchmark for reading comprehension with diverse types of test-time perturbation. These works focus on different domains than our research does, and they do not consider the composition of RE examples. Little attention is drawn to the entities in the sentences, and many attacks (e.g. character swapping, word injection) may make the perturbed sentences invalid. To the best of our knowledge, this work is among the first to propose a straightforward, dedicated pipeline for generating natural adversarial examples for the RE task, which takes into account the serious effects of entity bias in RE models.

6 Conclusion

Our contributions in this paper are three-fold. 1) Methodology-wise: we propose ENTRE, an end-to-end entity replacement method that reduces the shortcuts from entity names to ground-truth relations. 2) Resource-wise: we develop ENTRED, a straightforward method for generating natural and counterfactual entity replacements for RE, which produces ENTRED, a benchmark for auditing the generalization of RE models under entity bias. 3) Evaluation-wise: our experimental results and analysis provide answers to four main research questions on the generalization of RE. We believe ENTRED and the entity replacement method ENTRE can benefit the community working to increase the RE models' generalization under entity bias.

Acknowledgement

The authors would like to thank the anonymous reviewers for their discussion and feedback.

Wenxuan Zhou and Muhao Chen are supported by the NSF Grant IIS 2105329, the NSF Grant ITE 2333736, the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research, a Cisco Research Award, two Amazon Research Awards, and a Keston Research Award. Fei Wang is supported by the Annenberg Fellowship and the Amazon ML Fellowship. Yiwei Wang and Bryan Hooi are supported by NUS ODPRT Grant A-0008067-00-00, NUS ODPRT Grant R252-000-A81-133, and Singapore Ministry of Education Academic Research Fund Tier 3 under MOEs official grant number MOE2017-T3-1-007.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. 2015. Influence of resampling on accuracy of imbalanced classification. In *Eighth international conference on machine vision (ICMV 2015)*, volume 9875, page 987521. International Society for Optics and Photonics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Yi Yao Huang and William Yang Wang. 2017. Deep residual learning for weakly-supervised relation extraction. *arXiv preprint arXiv:1707.08866*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. 2020. Robust encodings: A framework for combating adversarial typos. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2752–2765, Online. Association for Computational Linguistics.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 65–74, Brussels, Belgium. Association for Computational Linguistics.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. *arXiv preprint arXiv:2109.05213*.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in MultiWOZ. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 326–337, Singapore and Online. Association for Computational Linguistics.

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

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2021. Benchmarking robustness of machine reading comprehension models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 634–644, Online. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the Thirty-fifth AAAI Conference on Aritificial Intelligence 2021*.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

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. *arXiv preprint arXiv:2205.03784*.

Nan Xu, Fei Wang, Bangzheng Li, Mingtao Dong, and Muhao Chen. 2022. Does your model classify entities reasonably? diagnosing and mitigating spurious correlations in entity typing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8658, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *arXiv preprint arXiv:2102.01373*.