

📄 RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models

Bill Yuchen Lin Wenyang Gao Jun Yan Ryan Moreno Xiang Ren
{yuchen.lin, wenyangg, yanjun, morenor, xiangren}@usc.edu
Department of Computer Science and Information Sciences Institute,
University of Southern California

Abstract

To audit the robustness of named entity recognition (NER) models, we propose RockNER, a simple yet effective method to create natural adversarial examples. Specifically, at the entity level, we replace target entities with other entities of the same semantic class in Wikidata; at the context level, we use pre-trained language models (e.g., BERT) to generate word substitutions. Together, the two levels of attack produce natural adversarial examples that result in a shifted distribution from the training data on which our target models have been trained. We apply the proposed method to the OntoNotes dataset and create a new benchmark named *OntoRock* for evaluating the robustness of existing NER models via a systematic evaluation protocol. Our experiments and analysis reveal that even the best model has a significant performance drop, and these models seem to memorize in-domain entity patterns instead of reasoning from the context. Our work also studies the effects of a few simple data augmentation methods to improve the robustness of NER models.¹

1 Introduction

Recent named entity recognition (NER) models have achieved great performance on many conventional benchmarks such as CoNLL2003 (Tjong Kim Sang, 2002) and OntoNotes 5.0 (Weischedel et al., 2013). However, it is not clear whether they are reliable in realistic applications in which entities and/or context words can be out of the distribution of the training data. It is thus important to audit the robustness of NER systems via natural adversarial attacks. Most existing methods for generating adversarial attacks in 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;

¹Our code and data are publicly available at the project website: <https://inklab.usc.edu/rockner>.

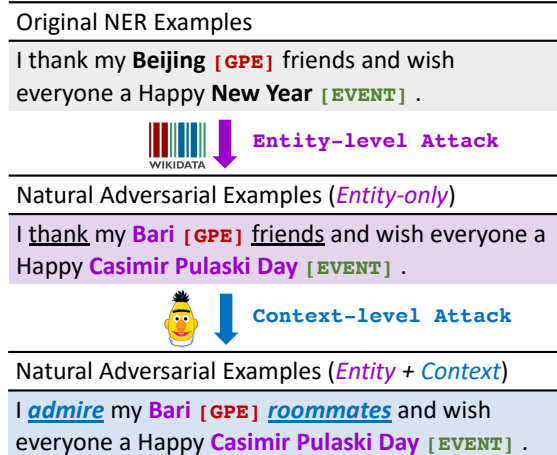


Figure 1: Illustration of RockNER attacking pipeline.

Gan and Ng, 2019), and these methods lack special designs reflecting the underlying compositions of the NER examples — i.e., *entity* structures and their *context* words. In this paper, we focus on creating general natural adversarial examples (i.e., real-world entities and human-readable context) for evaluating the robustness of NER models.

As shown in Figure 1, given a NER example, our method first generates entity-level attacks by replacing the original entities with entities from Wikidata and then uses a pre-trained masked language model like BERT (Devlin et al., 2019a) to generate context-level attacks. We choose the OntoNotes dataset (Weischedel et al., 2013)² to showcase ROCKNER because of the dataset’s high annotation quality and wide coverage of entity types. Thus, we create a novel benchmark, named *OntoRock*, for evaluating the robustness of a wide range of modern NER models.

We analyze the robustness of popular existing NER models on the *OntoRock* benchmark in order to answer three research questions as follows: (Q1) How robust are current NER models? (Q2)

²The proposed attacking method is also applicable to other general NER datasets.

Where are the NER models brittle? (Q3) Can we improve the robustness of NER models via data augmentation? Our experiments and analysis provide these main findings: 1) even the best model is still brittle to our natural adversarial examples, resulting in a significant performance drop (92.4% \rightarrow 58.5% in F1); 2) current NER models tend to memorize entity patterns instead of reasoning based on the context; also, there are specific patterns for entity typing mistakes; 3) simple data augmentation methods can indeed help us improve the robustness to some extent. We believe the proposed RockNER method, the OntoRock benchmark, and our analysis will benefit future research to improve the robustness of NER models.

2 Natural Adversarial Attacks for NER

We present RockNER, a simple yet effective method to generate high-quality natural adversarial examples for evaluating the robustness of NER models by perturbing both the entities and contexts of original examples. We apply the method to the development set and test set of OntoNotes to create the **OntoRock** benchmark.

2.1 Entity-Level Attacks

To generate relevant entities for modifying existing NER data, we collect a dictionary of natural adversarial entities of different fine-grained classes via Wikidata. As shown in Figure 2, our three-stage pipeline is introduced as follows:

- **(1) Entity Linking:** We first use *BLINK* (Wu et al., 2020) to link each entity in the original examples from its surface form to a canonical entry in Wikidata with a unique identifier (QID), e.g., “Beijing” \rightarrow *Q956*.
- **(2) Fine-grained Classification:** Then, we execute a query to get its fine-grained class via the *InstanceOf* relation (P31), e.g., *Q956* $\xrightarrow{P31}$ *Q1549591* (“big city”).
- **(3) Dictionary Expansion:** Finally, we retrieve additional Wikidata entities within each individual entity class. Given a particular entity such as “Beijing”, we collect additional *out-of-distribution* entities, such as “Bari”. They are both big cities (GPE type), while the latter one is *much less correlated* with the context in the training data³.

³We assure this by applying tested NER systems (§3) on sentences where each sentence is the an individual entity name, and only keep the ones the target models predict incorrectly.



Figure 2: Building adversarial entity dictionary.

To ensure the quality, we manually curate the fine-grained entity classes and remove entities linked incorrectly. We use a different approach for collecting PERSON attack entities because adversarial names can be more efficiently created as random combinations of first names, middle names, and last names, which are collected from the Wikidata person-name list⁴.

To create an evaluation benchmark based on existing datasets (e.g., OntoNotes), we iterate over every original entity and replace it with a randomly sampled adversarial entity from our dictionary sharing the *same fine-grained class*. We argue that the resulting attacks are both *natural* — i.e., containing real, valid entities, and *adversarial* — i.e., the entities are of the same class as the original entities while being out-of-distribution from the training data. Specifically, OntoRock has a much larger vocabulary of entity words than OntoNotes, and these words are rarely seen in the training set (Table 4 in Appendix §B). For example, for GPE and PRODUCT, OntoRock has $\sim 3x$ the number of unique entity words as OntoNotes has (GPE: 461 vs. 1202, PRODUCT: 54 vs. 158). The ratio of seen entity words is also much lower (GPE: 75.92% vs. 17.30%, PRODUCT: 44.44% vs. 7.59%).

2.2 Context-level Attacks

To investigate the robustness of NER models against changes to the context, we also create natural attacks on the context words. Our intuition

⁴We used the method described in <https://github.com/EdJoPaTo/wikidata-person-names>

	None	E	C	E+C
BLSTM-CRF	84.6	40.5 (↓ 52%)	77.3 (↓ 9%)	32.4 (↓ 62%)
SpaCy	87.3	43.9 (↓ 50%)	81.8 (↓ 6%)	40.1 (↓ 54%)
Stanza	87.9	56.1 (↓ 36%)	83.0 (↓ 6%)	51.7 (↓ 41%)
BERT-CRF	90.6	59.2 (↓ 35%)	85.8 (↓ 5%)	54.6 (↓ 40%)
Flair	90.7	59.6 (↓ 34%)	86.1 (↓ 5%)	55.3 (↓ 39%)
RoBERTa-CRF	92.4	63.4 (↓ 31%)	87.2 (↓ 6%)	58.5 (↓ 37%)
+ Ent. Switch.	91.4	64.7 (↓ 29%)	85.7 (↓ 6%)	59.1 (↓ 35%)
+ Rand. Mask.	92.6	66.3 (↓ 28%)	86.4 (↓ 7%)	60.0 (↓ 35%)
+ Mixing Up	92.0	61.1 (↓ 34%)	86.9 (↓ 6%)	56.5 (↓ 39%)

Table 1: F1 scores of models trained on OntoNotes’ training data and evaluated in different settings: **none** (original test of OntoNotes) and three variants of our OntoRock benchmark: (**E** for entity-only attacks, **C** for context-only attack, and **E+C** for the full version). Relative F1 drops shown as (↓ x).

is to replace context words with words that are semantically related and syntactically valid but out of the distribution of the training data. To this end, we perturb the original context by sampling adversarial tokens via a masked-language model such as BERT. Specifically, for each sentence, we choose semantic-rich words — nouns, verbs, adjectives, and adverbs — as the target tokens to replace. Then, we generate masked sentences with random numbers (at most 3) of [MASK] tokens. These masked sentences are then fed into BERT, which decodes the masked positions one by one from left to right. We use the predicted tokens ranking between 100~200, such that the words create more challenging context yet the sentence is still syntactically valid. As there are multiple sampled sentences, we take the one which is the least correlated with the training data. Specifically, we test all candidate sentences on the trained BLSTM-CRF model (which performs the worst among the target NER models) and we select the sentences that cause a performance drop.


2.3 OntoRock as a Robustness Benchmark

We create the most challenging version of our ROCKNER attack by applying both entity-level and context-level attacks on the original development and test sets of OntoNotes, forming our OntoRock benchmark. The overall statistics of OntoRock are shown in Table 3 (Appendix §B), alongside the statistics of the original OntoNotes dataset. We showcase RockNER using OntoNotes in this paper because of the dataset’s high annotation quality and comprehensive entity-type coverage. However, this method of attack is also appli-

cable to other datasets.

3 Evaluating Robustness of NER Models

In this section, we use our OntoRock dataset to evaluate the robustness of popular NER models including spaCy (Honnibal et al., 2020), Stanza (Qi et al., 2020), Flair (Akbik et al., 2018a), BLSTM-CRF (Lample et al., 2016), BERT-CRF (Devlin et al., 2019b), and RoBERTa-CRF (Liu et al., 2019). Model details are described in Appendix §C. We organize our results and analysis as three main research questions and their answers.

 **Q1: How robust are current NER models?**

Main results. We show the F1 scores on the test sets⁵ in Table 1. We can see that all NER models have a significant performance drop in the attacked settings (i.e., entity attack only, context attack only, and both); there is a 35% ~ 62% relative decrease (in the models’ F1) in the fully-attacked setting as compared to their results⁶ on the original test set. We find the performance on the original test set is positively correlated with the robustness against our attacks. Thus, models that perform better on in-domain data tend to be also better at handling out-of-distribution examples.

Pre-training & Robustness. BLSTM-CRF is trained solely on the training set of OntoNotes; The NER toolkits such as spaCy and Stanza are trained on more datasets (e.g., CoNLL03); BERT-CRF, Flair, and RoBERTa-CRF are based on pre-trained language models. We can see that, in terms of robustness, NER models with pre-training tend to outperform models without pre-training but with more NER data access, which outperform those trained only on the OntoNotes training set. This observation indicates that pre-training (on corpora or other NER data) leads to better robustness, and better pre-trained models (RoBERTa vs. BERT) have a lower (relative) performance drop. Interestingly, we find that the improved robustness from pre-training mainly comes from the improvement on the entity-level attacks, possibly because of the increased exposure to entities and increased ability to reason using context (see our 1st point in Q2).

 **Q2: Where are the NER models brittle?**

⁵Full results on dev and test are reported in Appendix §D.

⁶All models are trained on the OntoNotes’ training data.

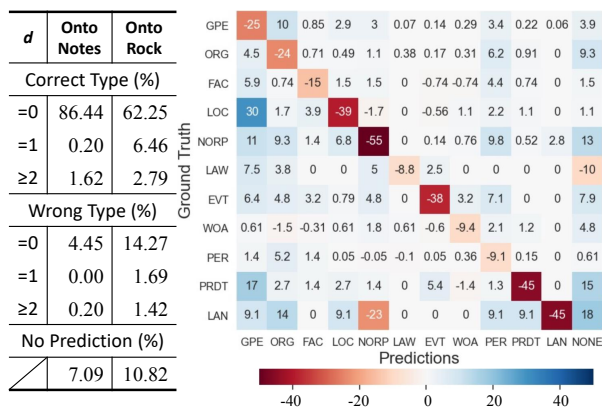


Figure 3: Error analysis of RoBERTa-CRF. **Left:** Difference of predictions on OntoNotes and OntoRock. **Right:** Difference of confusion matrices.

Memorizing or Reasoning? Note that our entity-level attacks aim to test the ability to use the context to infer the entities, as the novel entities themselves are out-of-distribution — i.e., if a model can reason about the context, it should be robust against entity changes. In turn, the context-level attacks audit the ability to memorize entity patterns, as the context is changed, making it more challenging to infer from. From Table 1, we can see that all models have a smaller performance drop in context-level attacks and a larger performance drop in entity-level attacks. Therefore, we conclude that NER models are apt to memorize entity patterns presented in the training data and are more brittle when emerging, out-of-distribution entities exist in the inputs. This also suggests that current NER models tend to infer the type and boundary of entities without properly using the context. To make NER models more robust, we believe an important future direction is to develop context-based reasoning approaches, taking advantage of inductive biases such as entity triggers (Lin et al., 2020).

Error Analysis. To analyze the additional errors caused by our attacks, we look at each truth entity and inspect the changes of model behaviors in this position. We pair each original entity with its overlapped prediction and categorize it as follows: (1) whether the predicted *type* matches (Correct/Wrong); (2) the number of *different tokens* between the prediction and truth (d). In Figure 3 (left), RoBERTa-CRF’s predictions on OntoNotes and OntoRock and find that most additional error cases (86.4% vs. 62.3%) are caused by typing errors — the model either predicts a wrong type (4.5% vs. 14.3%), or NONE (7.1%

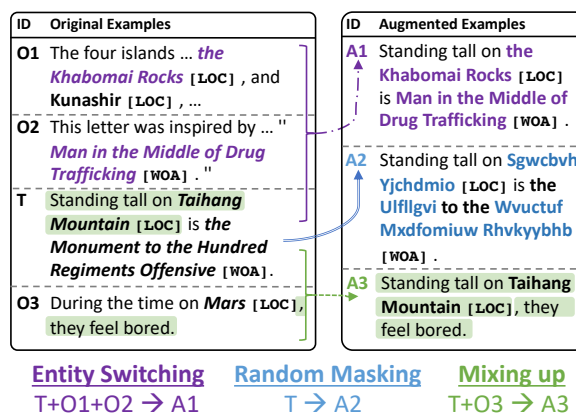


Figure 4: Three simple methods for augmenting training data to improve the robustness of NER.

vs. 10.8%). Concrete cases are shown in Figure 5 (Appendix §E).

We take a closer look by calculating the difference between the models’ confusion matrices on the attacked and the original test data (i.e., OntoRock’s minus OntoNotes’), as shown in Figure 3 (right). This confusion-difference matrix reveals the model’s weakness in handling novel entities, especially when making decisions between closely related categories. For example, the biggest difference is the typing error from LOC to GPE (increased by 30 points)⁷ — indicating that the model struggles to recognize names of countries/cities/states that are not covered by the distribution of training data.

Apart from that, we find that the entity-level and context-level attacks succeed in different parts of examples. We denote the sets of entity spans that are mistakenly predicted in entity-only attacks and context-only attacks as S_E and S_C . Their Jaccard similarity is only 0.04, which shows that these two attacks target different kinds of weaknesses.

Q3: Can we improve the robustness of NER models via data augmentation?

Methods. The most straightforward method to improve NER robustness is to augment our examples used for training models. Here we use three intuitive data augmentation methods for the analysis. **1) Entity Switching:** we replace each entity in the target sentence with a different entity of the same type from another sentence. **2) Random Masking:** for each entity, we replace every one of its letters with a random one. We

⁷GPE (Geo-Political Entity) is defined to include “countries, cities, states”, while LOC (Location) is defined as “non-GPE locations, mountain ranges, bodies of water”.

retain the same capitalization pattern and keep all stopwords unchanged. **3) Mixing up:** inspired by Guo et al. (2019), we randomly pick one entity from the target sentence and find another sentence that includes an entity of the same type; then we generate an adversarial sentence by merging the first half of the target sentence (up to and including the entity) with the second half of the second sentence (everything after the entity). They are illustrated with examples in Figure 4.

Results & Analysis. The results of the three methods on the RoBERTa-CRF model are shown in Table 1. Surprisingly, the most straightforward method, Random Masking, offers the best improvement against entity-level attacks. We conjecture it is because it provides more entity patterns, which enhances its entity-level generalization ability and makes models focus more on the context for inference, resulting in a better performance on entity-level attacks (63.4% \rightarrow 66.3%). As the Entity Switching repeats original entities in the different context of the training set, it aims to improve the performance in using context to infer entities. The entity-level attacks are indeed better handled (63.4% \rightarrow 64.7%). The Mixing up method, however, loses the robustness on all settings, possibly due to potential noise from sentences that are not syntactically valid.

4 Related Work

There are other recent works which also turn their attention from achieving a new state-of-the-art of NER model towards studying NER models’ robustness and generalization ability. Agarwal et al. (2020a) create entity-switched datasets by replacing entities with others of the same type but different national origin. They find that NER models perform worse on entities from certain countries. Mayhew et al. (2020) and Bodapati et al. (2019) focus on the robustness when inputs are not written in the standard casing (e.g., “he is from us” \rightarrow “US”). Fu et al. (2020) analyze the generalization ability of current NER models by evaluating them across datasets. Agarwal et al. (2020b) further analyze the roles of context and names in entity predictions made by models and humans. Although these works begin to understand the robustness issue of NER models, they do not build an automated pipeline to generate natural adversarial instances with large coverage (e.g. thousands of fine-grained classes) at scale.

There are also works in other domains aiming to evaluate models’ robustness with perturbed inputs. For example, Jia and Liang (2017) attack 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’ over-sensitivity. 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 NER 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 NER task, which takes into account the compositions of NER examples — *i.e.*, entity structures and their context.

5 Conclusion

Our contributions in this short paper are two-fold. 1) resource-wise: we develop RockNER, a straightforward method for generating natural adversarial attacks for NER, which produces **OntoRock**, a benchmark for auditing the robustness of NER models. 2) evaluation-wise: our experimental results and analysis provides answers supported by experimental results to three main research questions on the robustness of current mainstream NER models. We believe *RockNER* and its produced attacks (e.g., the *OntoRock* benchmark) can benefit the community working to increase the robustness and out-of-distribution generalization of NER.⁸

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007, the DARPA MCS program under Contract No. N660011924033, the Defense Advanced Research Projects Agency with award W911NF-19-20271, NSF IIS 2048211, NSF SMA 1829268, and gift awards from Google, Amazon, JP Morgan and Sony. We would like to thank the reviewers for their constructive feedback.

⁸We leave our full results, more implementation details, and additional analyses in the appendix.

References

- Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and A. Nenkova. 2020a. [Entity-switched datasets: An approach to auditing the in-domain robustness of named entity recognition models](#). *ArXiv preprint*, abs/2004.04123.
- Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and A. Nenkova. 2020b. [Interpretability analysis for named entity recognition to understand system predictions and how they can improve](#). *ArXiv preprint*, abs/2004.04564.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018a. [Contextual string embeddings for sequence labeling](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018b. [Contextual string embeddings for sequence labeling](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Pravrajit Bodapati, Hyokun Yun, and Yaser Al-Onaizan. 2019. [Robustness to capitalization errors in named entity recognition](#). In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 237–242, Hong Kong, China. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jinlan Fu, Pengfei Liu, and Qi Zhang. 2020. [Rethinking generalization of neural models: A named entity recognition case study](#). 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 7732–7739. AAAI Press.
- 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.
- Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. [Augmenting data with mixup for sentence classification: An empirical study](#). *ArXiv preprint*, abs/1905.08941.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- 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.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. 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.
- Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. 2020. [TriggerNER: Learning with entity triggers as explanations for named entity recognition](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8503–8511, Online. 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*, abs/1907.11692.
- Stephen Mayhew, Nitish Gupta, and Dan Roth. 2020. [Robust named entity recognition with truecasing pre-training](#). 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 8480–8487. AAAI Press.
- 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.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. [Towards robust linguistic analysis using OntoNotes](#). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, 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.
- Erik F. Tjong Kim Sang. 2002. [Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition](#). In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.
- Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. [Scalable zero-shot entity linking with dense entity retrieval](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

A Statistics of the Entity Dictionary

In Table 2, we show statistics of the adversarial entity dictionary built for the test set. We generate 279,290 adversarial entities out of 7,433 original entities. The amount of generated entities is 36 times larger than the original ones, which extremely enriches candidates for conducting entity-level attacks. Moreover, there is one class for every 2~10 entities according to their type, and each class includes hundreds of adversarial entities. This indicates that we have enough adversarial entities to conduct entity-level attacks.

Type	# Original	# Classes	# Adversarial
GPE	2,203	237	42,912
ORG	1,750	402	75,259
FAC	135	84	29,041
LOC	179	75	20,354
NORP	830	117	27,033
LAW	40	14	3,920
EVENT	63	33	9,636
WOA	165	73	39,508
PRODUCT	74	48	17,986
LANG	22	7	4,119
PERSON	1,972	N/A	9,522
Total	7,433	1,090	279,290

Table 2: Statistics of the adversarial entity dictionary.

B Statistics of the Dataset

	Train	Dev	Test
# Sentences	59,924	8,528	8,262
# Tokens	1.1M	148k	153k
# Entities	55,008	7,482	7,433
# Attacked Entities	N/A	6,962	6,939
% Attacked Entities	N/A	93.05	93.35
# Attacked Context Words	N/A	16,155	15,664
% Attacked Sentences	N/A	98.03	97.53

Table 3: Overall statistics of OntoRock benchmark.

We adopt the train/dev/test splits of OntoNotes used by Pradhan et al. (2013) in our experiments. Table 3 presents the statistics of our OntoRock benchmark, which consists of the original OntoNotes training set and our attacked (full version) development and test sets. Table 4 shows the statistics of entities in the training set, OntoNotes’ test set and OntoRock’s test set.

	Train	OntoNotes Test		OntoRock Test	
	# ent_words	# ent_words	seen (%)	# ent_words	seen (%)
GPE	1615	461	75.92	1202	17.30
ORG	4037	1056	67.42	2399	26.18
FAC	681	125	55.20	287	19.16
LOC	527	118	66.95	224	22.32
NORP	565	160	78.13	330	8.18
LAW	343	74	45.95	116	26.72
EVENT	439	91	64.84	132	23.48
WOA	1107	264	37.12	351	17.95
PERSON	5367	1102	61.62	1011	31.95
PRODUCT	381	54	44.44	158	7.59
LANGUAGE	33	7	71.43	25	4.00
ALL_TYPES	12174	3028	68.13	5522	31.44

Table 4: Entity statistics of OntoNotes and OntoRock benchmarks.

C Model Details

For spaCy, we load the “en_core_web_lg” model with the white-space tokenizer.

For the Stanza model, we use the English model and set processors as “tokenize, ner” with tokenize_pretokenized= True.

When we train the Flair model with a GPU, we set mini_batch_size as 64, train_with_dev as False and embeddings_storage_mode as “none”.

For training BLSTM-CRF, BERT-CRF and RoBERTa-CRF models, we set batch_size as 20. We use early stopping and set patience=10 for BLSTM-CRF and 5 for the other two.

D Full Results

Precision/Recall/F1 scores for each model on the original OntoNotes and our OntoRock benchmark are presented in Table 6 (test set) and Table 7 (development set).

E Cases

In Fig. 5, we show examples of entity-level attacks on the RoBERTa. These examples should be easily solved based on the context. For example, "a host of" in sentence 1 and "holiday" in sentence 4 are both explicit clues. If NER models are capable of inferring from context, those clues could have assisted them to achieve better performance. It qualitatively validates our hypothesis that NER models tend to remember entity patterns instead of inferring entity labels from context.

F More Details of Error Analysis

In Figure 7, we present the confusion matrices for RoBERTa-CRF model on the OntoNotes’ and OntoRock’s test sets. We use them to calculate the confusion difference matrix (Figure 3 (right)).

Robustness Analysis (by prediction)		RoBERTa-CRF (%)		RB-CRF+ES (%)		RB-CRF+RM (%)		RB-CRF+M (%)	
		Attacked	Unattacked	Attacked	Unattacked	Attacked	Unattacked	Attacked	Unattacked
Correct Type	d=0 (SameSpan)	62.55	86.44	64.78	83.40	66.41	86.23	60.59	84.41
	d=1	6.46	0.20	5.76	0.00	6.49	0.20	7.09	1.42
	d=2	1.02	0.61	0.60	1.01	1.07	0.81	0.73	1.01
	d≥3	1.77	1.01	1.69	2.53	1.75	1.21	1.51	1.42
Wrong Type	d=0 (SameSpan)	14.27	4.45	14.20	4.45	12.88	4.25	14.43	4.25
	d=1	1.69	0.00	1.50	0.00	1.59	0.00	1.77	0.00
	d=2	0.59	0.20	0.72	0.00	0.55	0.20	0.59	0.20
	d≥3	0.83	0.00	1.19	0.10	1.15	0.20	1.28	0.20
No Prediction		10.82	7.09	9.57	8.50	8.10	6.88	12.02	7.09

Table 5: Error analysis of RoBERTa-CRF and with three different data augmentation methods. d indicates the number of different indices between ground-truth entity and the overlapped predictions (**ES** for Entity Switching, **RM** for Random Masking, and **M** for Mixing up).

ID	Original Sentence	Attacked Sentence	Sentence with Predicted Tags
1	Next is Yang Yang [PERSON], a host of Beijing Traffic Radio Station [ORG].	Next is A.A. Sidhu [PERSON], a host of Beijing Tiyu Guangbo [ORG].	Next is A.A. Sidhu [PERSON], a host of Beijing Tiyu Guangbo [PERSON].
2	It might even impact traffic on the Second Ring Road [FAC] and Fourth Ring Road [FAC].	It might even impact traffic on the R15 road [FAC] and A821 autoroute [FAC].	It might even impact traffic on the R15 road and A821 autoroute [FAC].
3	A friend said, I work on the 12th floor of the China World Trade Center [FAC] at the Guomao Bridge [FAC], ah.	A friend said, I work on the 12th floor of Jim Henson Company Lot [FAC] at the Guomao Bridge [FAC], ah.	A friend said, I work on the 12th floor of Jim Henson Company [ORG] Lot at the Guomao Bridge [FAC], ah.
4	Ah, today is the first workday after the New Year [EVENT] holiday.	Ah, today is the first workday after the Restoration Day of the Independent Czech State [EVENT] holiday.	Ah, today is the first workday after the Restoration Day of the Independent Czech [NORP] State holiday.
5	I think, in comparison to China [GPE], we should say that urbanization in foreign countries developed earlier and is more widespread.	I think, in comparison to Danish Realm [GPE], we should say that urbanization in foreign countries developed earlier and is more widespread.	I think, in comparison to Danish [NORP] Realm, we should say that urbanization in foreign countries developed earlier and is more widespread.

Figure 5: Example cases for the entity-level attacks.

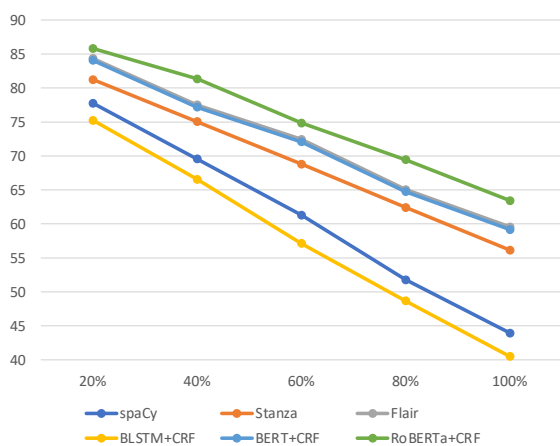


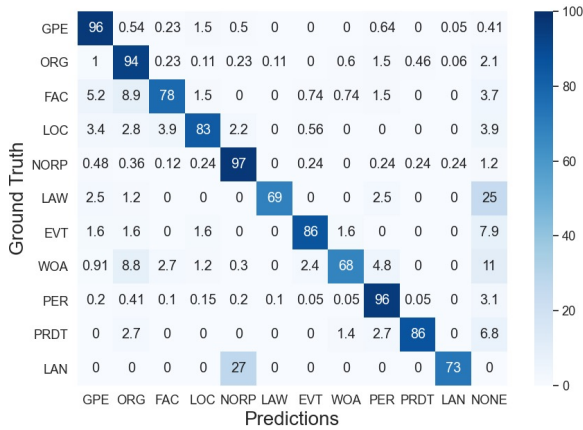
Figure 6: F1 scores for the RoBERTa-CRF model under entity-level attacks with different coverage.

Following Figure 3 (left), we categorize the error cases of predictions by the RoBERTa-CRF model and its variants that are trained with augmented data. The results are presented in Table 5. Among three augmentation methods, random masking gets the highest F1 score on both attacked and original test sets. The robustness gain mainly comes from

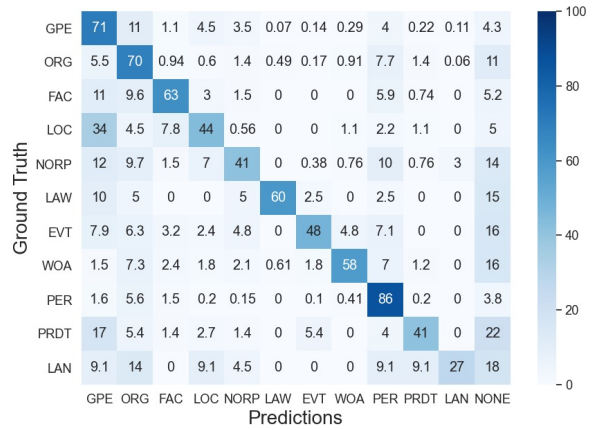
more accurate typing (Wrong Type, $d = 0$: 14.27% \rightarrow 12.88%).

G Attacking Curve

For the entity-level attacks, we conduct 5 separate attacks by replacing 20%, 40%, 60%, 80%, and 100% of the entities in the test set. For each model, we evaluate it on the 5 generated test sets and plot a curve of the F1 scores for each attack, shown in Fig. 6. The descending trend is intuitive, and the performance of weak models drops more rapidly than the performance of strong models.



(a) Confusion matrix for ground truth and predictions of RoBERTa-CRF model on the OntoNotes test set.



(b) Confusion matrix for ground truth and predictions of RoBERTa-CRF model on the OntoRock test set.

Figure 7: Confusion matrices for RoBERTa-CRF on OntoNotes’ and OntoRock’s test sets.

Evaluation Metrics →	Precision (%)				Recall (%)				F1 (%)			
	None	E	C	E+C	None	E	C	E+C	None	E	C	E+C
BLSTM-CRF (Lample et al., 2016)	86.3	40.8	78.7	32.7	82.9	40.2	76.1	32.0	84.6	40.5	77.3	32.4
spaCy (Honnibal et al., 2020)	88.3	43.9	82.0	40.1	86.2	44.0	81.6	40.1	87.3	43.9	81.8	40.1
Stanza (Qi et al., 2020)	89.5	55.3	83.9	51.2	86.3	57.1	82.1	52.2	87.9	56.1	83.0	51.7
BERT-CRF (Devlin et al., 2019a)	91.9	59.8	86.7	55.7	89.3	58.5	84.9	53.5	90.6	59.2	85.8	54.6
Flair (Akbik et al., 2018b)	91.3	59.4	86.0	55.3	90.2	59.8	86.1	55.2	90.7	59.6	86.1	55.3
RoBERTa-CRF (Liu et al., 2019)	92.9	63.1	87.4	58.5	91.8	63.7	87.0	58.5	92.4	63.4	87.2	58.5
RB-CRF+Entity Switching	92.2	64.2	85.7	58.7	90.6	65.2	85.7	59.5	91.4	64.7	85.7	59.1
RB-CRF+Random Masking	92.8	65.3	86.1	59.2	92.4	67.3	86.8	60.8	92.6	66.3	86.4	60.0
RB-CRF+Mixing Up	92.5	60.7	86.7	56.3	91.4	61.5	87.0	56.8	92.0	61.1	86.9	56.5

Table 6: Results of NER models on the test set of OntoNotes with none changes and three variants of the OntoRock benchmark (E for entity-only attacks, C for context-only attacks, and E+C for the full version).

Evaluation Metrics →	Precision (%)				Recall (%)				F1 (%)			
	None	E	C	E+C	None	E	C	E+C	None	E	C	E+C
BLSTM-CRF (Lample et al., 2016)	85.4	40.0	77.2	33.2	82.5	39.9	75.2	32.6	83.9	40.0	76.2	32.9
spaCy (Honnibal et al., 2020)	86.0	43.2	79.7	40.3	84.8	44.0	79.7	40.4	85.4	43.6	79.7	40.3
Stanza (Qi et al., 2020)	87.5	52.6	81.1	48.8	84.2	55.2	79.2	50.8	85.8	53.9	80.1	49.8
BERT-CRF (Devlin et al., 2019a)	91.2	58.1	84.8	54.3	88.9	57.3	83.1	52.7	90.0	57.7	84.0	53.5
Flair (Akbik et al., 2018b)	89.4	56.6	83.6	52.5	89.0	58.1	84.1	53.2	89.2	57.3	83.9	52.9
RoBERTa-CRF (Liu et al., 2019)	91.3	60.9	84.9	56.1	90.4	62.2	84.8	56.9	90.0	61.6	84.8	56.5
RB-CRF+Entity Switching	90.3	61.8	83.5	56.1	89.5	64.5	83.8	58.1	89.9	63.1	83.7	57.1
RB-CRF+Random Masking	90.6	62.5	83.7	56.4	90.7	65.2	84.9	58.5	90.7	63.8	84.3	57.4
RB-CRF+Mixing Up	90.8	58.1	84.4	53.5	90.6	60.0	85.5	55.2	90.7	59.0	85.0	54.4

Table 7: Results of NER models on the development set of OntoNotes with none changes and three variants of the OntoRock benchmark (E for entity-only attacks, C for context-only attacks, and E+C for the full version).