Resilience of Named Entity Recognition Models Under Adversarial Attack

Sudeshna Das
Indian Institute of Technology
Kharagpur
sudeshna.das@iitkgp.ac.in

Jiaul H Paik
Indian Institute of Technology
Kharagpur
jiaul@cet.iitkgp.ac.in

Abstract

Named entity recognition (NER) is a popular language processing task with wide applications. Progress in NER has been noteworthy, as evidenced by the F1 scores obtained on standard datasets. In practice, however, the end-user uses an NER model on their dataset out-of-the-box, on text that may not be pristine. In this paper we present four model-agnostic adversarial attacks to gauge the resilience of NER models in such scenarios. Our experiments on four state-of-the-art NER methods with five English datasets suggest that the NER models are over-reliant on case information and do not utilise contextual information well. As such, they are highly susceptible to adversarial attacks based on these features.

1 Introduction

Named entity recognition (NER) is a popular language processing task that involves identifying and classifying named entities in text (Mayhew et al., 2020). Progress in NER has been rapid and noteworthy, especially in the current age of deep learning (Li et al., 2020). The general impetus in deep learning-based NER has been to develop models that incorporate context better (Akbik et al., 2018; Devlin et al., 2019; Manning et al., 2014) and are resilient to noise such as inconsistencies in case information (Mayhew et al., 2019; Bodapati et al., 2019; Mayhew et al., 2020). There has, however, been modest focus on determining the extent to which state-of-the-art NER models succeed in doing so. Identifying the weaknesses of NER models can help drive focused work to ameliorate them and move NER beyond marginal improvements in F1 scores (Stanislawek et al., 2019).

Adversarial attacks designed for NLP models largely focus on classification tasks (Wallace et al., 2019; Ren et al., 2019; Jia et al., 2019; Wallace et al., 2019; Papernot et al., 2016). Many existing studies work with vector representations (Ebrahimi et al., 2018; Zhao et al., 2018), which are not intuitively interpretable by humans. Such methods require white-box access to the models (Ren et al., 2019). The additional requirement of human intervention to adjudge the quality of adversarial samples generated may also be involved (Alzantot et al., 2018).

Adversarial NER has broadly seen two types of approaches: (a) adversarial training, and (b) adversarial evaluation. Adversarial training of NER models involves introducing small perturbations in the training data to make models robust (Bekoulis et al., 2018). Such perturbations are introduced in the text representation level (Wang et al., 2020; Bai et al., 2020; Huang et al., 2022). The adversarial evaluation of NER models, on the other hand, involves benchmarking the models on synthetically generated data (Lin et al., 2021; Simoncini and Spanakis, 2021). We follow the latter line of investigation.

We present four model-agnostic adversarial attacks targeted at NER models. Our task-specific approach allows us to generate natural language adversaries that work with pre-trained models and are easily interpretable by humans. In principle, our work is similar to the label-preserving substitutions explored by Ren et al. (2019) and the word-substitution methods explored by Alzantot et al. (2018), although they do not evaluate their methods on NER. Generating adversarial data for evaluating NER models is explored by Simoncini and Spanakis (2021) using BERT to replace and/or add non-named entity tokens to text. Lin et al. (2021) also use pre-trained BERT to generate context-level adversarial attacks to evaluate NER models. In contrast to their work, we use simple rule-based methods for generating adversarial data. Our method has the advantage of not requiring re-training or fine-tuning of pre-trained models.

The datasets and models we use are all openly available, aiding reproducibility.[1] Further, our ex-

[1] https://github.com/das-sudeshna/adversarial-ner

Proceedings of the First Workshop on Dynamic Adversarial Data Collection, pages 1 - 6

	CoNLL	WIKI	GMB	FIRE	IEER
# LOCATION	1668	1014	59255	2626	878
# PERSON	1617	934	18970	2725	1504
# ORGANIZATION	1661	898	22662	893	939

Table 1: Data description: frequency of named entities.

periments do not require white-box access to the models.

2 Data

We use five openly available general domain datasets that contain the *enamex* classes (LOCATION, PERSON, & ORGANIZATION) (Nadeau and Sekine, 2007), for this study.

CoNLL-2003 The CoNLL-2003 (CoNLL) dataset consists of news articles from the Reuters Corpus (Tjong Kim Sang and De Meulder, 2003). In keeping with the standard evaluation schemes, we report results only on the test split of the dataset.

WikiGold The WikiGold dataset (WIKI) comprises of manually annotated English Wikipedia articles (Balasuriya et al., 2009).

FIRE NER 2013 The English dataset (FIRE) from the *NER for Indian Languages* task at FIRE 2013 comprises of text crawled from Indian websites as well as Wikipedia articles.

NIST IE-ER 1999 IEER refers to the gold standard NEWSWIRE development test data for the NIST 1999 IE-ER Evaluation available with NLTK (Steven Bird and Klein, 2009).

GMB 2.2 The Groningen Meaning Bank 2.2 dataset comprises of public domain texts that include news articles, stories, jokes, and transcripts. NLP tools are used to provide a preliminary annotation which is then updated by a combination of human experts, NLP tools, and crowd-sourcing to yield a silver-standard corpus (Bos et al., 2017).

3 Methods

We use four named entity recognizers for our experiments, all of which are open-source. Of these, spaCy is the current state-of-the-art in terms of document processing speed (Choi et al., 2015) and Flair is near the current state-of-the-art.[2]

[2]The F1 score of the current state-of-the-art model is 0.935. (Flair's F1 score is 0.931.) Since a pre-trained model is not publicly available, we choose not to include it in our experiments. We strongly believe that this does not affect the conclusions of our work.

Flair NER The Flair named entity recognizer is based on neural character embeddings. It uses contextual neural string embeddings that are obtained by pre-training on large, unlabelled corpora. Every sentence is represented in the form of string embeddings which are then stacked with pre-computed uncased GloVe embeddings, before being passed through a BiLSTM-CRF architecture that generates labels for each word (Akbik et al., 2018).

spaCy NER spaCy's named entity recognizer employs a transition-based entity recognition methodology where state changes are triggered by actions. It uses trigram CNNs with residual connections that transform context-independent vectors into context-sensitive vectors (Honnibal, 2016).

CoreNLP NER CoreNLP NER (Manning et al., 2014) is based on linear chain Conditional Random Field (CRF) sequence models of arbitrary order (Finkel et al., 2005). For our experiments, we use the caseless model that ignores capitalization as well as the Truecase annotator that attempts to rectify incorrect casing, in addition to the default model.

DeepPavlov NER DeepPavlov's named entity recognition model uses the English cased model of BERT with 12 layers, 768 hidden nodes, 12 attention heads, and 110M parameters (Devlin et al., 2019). The first sub-word representation of each word is passed through a dense layer to generate labels (Burtsev et al., 2018).

Original:	My sister Alice (PERSON) lives in London (LOCATION).
Case	
Ablation:	my sister alice (PERSON) lives in london (LOCATION).
Aberration:	My sister alice (PERSON) Lives in London (LOCATION).
Context	
Perturbation:	My sister London (PERSON) lives in Alice (LOCATION).
Alteration:	My sister lives Alice (PERSON*) London (LOCATION*) in.

Figure 1: Dataset variants. Class* denotes named entities that should desirably be misclassified from their context.

4 Adversarial Attacks

In this section we describe the design of two broad types of adversarial attacks on NER models.

4.1 Case-based Adversarial Attacks

Case is one of the strongest indicators of named entities in English (Mayhew et al., 2020) and it is well known that case affects the performance of NER models (Mayhew et al., 2019; Bodapati et al., 2019). We formulate two adversarial attacks that emulate data where (i) case information may be unavailable, such as informal texts, and (ii) case information is unreliable, such as text extracted from PDF or OCR-ed documents.

4.1.1 Case Ablation

In case ablation, we drop the case information while keeping the rest of the text intact. The case-ablated named entities attempt to fool the NER models into misclassifying them as non-entities. This allows us to quantify what percentage of the correctly identified named entities rely completely on case information.

4.1.2 Case Aberration

In this setup, we randomly capitalise N percent of the tokens in each dataset, where N is the percentage of actual named entity tokens in the corresponding original text. The randomly capitalised tokens attempt to fool the model into marking them as named entities. We choose N rather than an arbitrary value in order to maintain the distribution of capitalised and lowercase tokens in the datasets.

4.2 Context-based Adversarial Attacks

The surrounding text of a named entity is arguably the most useful feature in identifying named entities. All the NER models we evaluate attempt to capture context to leverage this information. We formulate two adversarial attacks that attempt to determine how well such information is captured by these models.

4.2.1 Context Perturbation

We create local perturbations for named entities. That is, we change the immediately surrounding text of the named entities while retaining syntactic structure and a semblance of semantics. To achieve this, we replace named entities of each class by named entities of the other two classes, with an equal probability. The local context of a named entity attempts to fool the NER model into classifying it incorrectly. This attack is similar in nature to the data augmentation procedure used by Lin et al. (2021). However, they restrict named entity substitutions within the same entity class. Since we carry out inter-class entity substitutions, we posit that our method is better able to detect when NER models rely on memorising named entity tokens.

4.2.2 Context Alteration

We alter the context of named entities on a global scale. To achieve this, we randomly select named entities with equal probability and place them in random locations in the text. In almost all cases, the text becomes grammatically incorrect, as is illustrated in Table 1. Thus, neither semantics nor syntactic rules are maintained, effectively altering the global contextual frame of named entities. In this case, it is *desirable* for models to misclassify named entities. That is, we consider a model to be better if it is *susceptible* to this attack. This is based on our hypothesis that a model that captures context better should perform *worse* when the context is meaningless.

5 Evaluation

We follow the CoNLL-2003 Shared Task guidelines to report the F1 scores (Tjong Kim Sang and De Meulder, 2003). Compatible classes are clubbed with the closest enamex class (such as, GPE (Geo-political entity) is clubbed with LOCATION for spaCy, BERT, and the GMB dataset). The class labels present in different datasets/produced by different models do not always have a close one-to-one correspondence to the class labels in other datasets/produced by other models. Thus, non-enamex entities are considered to be non-entities to provide a fair comparison across datasets and models. NER models and datasets also differ in their tagging schemes. Since it is not possible to map IO tags to IOB or IOBES, and IOB tags to IOBES (Cho et al., 2013), we map all tags into the IO scheme. The mapping of compatible entity classes and tagging schemes causes our evaluation results to differ from the officially reported scores of these NER models.

Model	CoNLL	WIKI	GMB	FIRE	IEER
Flair	**0.92**	**0.92**	**0.93**	**0.93**	**0.93**
spaCy	0.85	0.89	0.91	0.90	0.90
CoreNLP-s	0.86	0.89	0.92	0.88	0.90
DeepPavlov	0.83	0.90	0.90	0.90	0.92

Table 2: F1 scores on original datasets.

6 Results and Analysis

Table 2 shows the F1 scores of the models on the original dataset. This gives us the benchmark against which we compare the performance for the different data variants.

Model	CoNLL	WIKI	GMB	FIRE	IEER
Flair	**0.37** (-52.03%)	0.16 (-74.24%)	0.35 (-54.48%)	0.18 (-73.66%)	0.14 (-77.97%)
spaCy	0.14 (-68.64%)	0.13 (-74.06%)	0.20 (-68.32%)	0.15 (-72.97%)	0.14 (-75.02%)
CoreNLP-s	0.19 (-62.81%)	0.11 (-75.73%)	0.28 (-61.76%)	0.16 (-69.57%)	0.12 (-77.11%)
CoreNLP-c	0.32 **(-47.72%)**	**0.23 (-62.75%)**	**0.42 (-46.74%)**	**0.30 (-53.83%)**	**0.28 (-59.58%)**
CoreNLP-t	0.20 (-62.69%)	0.11 (-75.96%)	0.28 (-61.76%)	0.16 (-70.03%)	0.11 (-77.78%)
DeepPavlov	0.25 (-52.45%)	0.15 (-73.30%)	0.16 (-72.42%)	0.23 (-63.91%)	0.13 (-77.77%)

Table 3: F1 scores on case ablated datasets. High F1 score and low percentage drops are desirable.

Model	CoNLL	WIKI	GMB	FIRE	IEER
Flair	**0.39** (-50.19%)	0.21 (-69.00%)	0.36 (-53.62%)	0.23 (-68.50%)	0.21 (-71.33%)
spaCy	0.20 (-61.21%)	0.19 (-67.43%)	0.26 (-61.70%)	0.20 (-67.61%)	0.19 (-69.82%)
CoreNLP-s	0.25 (-56.38%)	0.17 (-69.19%)	0.32 (-57.08%)	0.21 (-63.75%)	0.17 (-71.34%)
CoreNLP-c	0.32 **(-47.72%)**	**0.23 (-62.75%)**	**0.42 (-46.74%)**	**0.30 (-53.83%)**	**0.28 (-59.58%)**
CoreNLP-t	0.25 (-56.38%)	0.17 (-69.19%)	0.32 (-57.08%)	0.20 (-64.55%)	0.16 (-72.12%)
DeepPavlov	0.32 (-44.46%)	0.15 (-72.29%)	0.22 (-65.51%)	0.26 (-60.23%)	0.19 (-71.59%)

Table 4: F1 scores on case aberrated datasets. High F1 score and low percentage drops are desirable.

6.1 Case ablation

We observe significantly large performance drops for every model with respect to model performance on the original datasets. This is unsurprising, as case information is an important indicator of named entities.

If we consider the CoreNLP-c scores as the upper bound (since this model is trained on caseless data and hence, reflects the ability of NER models to work on caseless data), we still notice large drops in F1 scores for the other models. This reflects the tendency of NER models to over-rely on case information. Among the cased models, we find BERT to be the better performer with Flair trailing as a close competitor. This is an interesting finding as it suggests that cased BERT is more resilient to case-based adversarial attacks than Flair, which uses uncased GloVe embeddings.

6.2 Case aberration

We observe large drops in performance for the case aberration attack. The performance for CoreNLP-t is worse than that of CoreNLP-c, which suggests that truecasing is not as effective as caseless training. Among the case-sensitive models, we find Flair outperforming other models. The performance drop for case aberration is slightly less than that for case ablation.

Model	CoNLL	WIKI	GMB	FIRE	IEER
Flair	**0.22** (-68.53%)	**0.27 (-62.05%)**	**0.23 (-67.65%)**	**0.19 (-73.05%)**	**0.18 (-74.22%)**
spaCy	0.15 (-67.49%)	0.20 (-66.96%)	0.19 (-69.74%)	0.13 (-74.71%)	0.12 (-77.03%)
CoreNLP-s	0.16 **(-67.02%)**	0.17 (-68.99%)	0.19 (-71.52%)	0.13 (-73.37%)	0.09 (-79.73%)
DeepPavlov	0.11 (-69.28%)	0.10 (-78.33%)	0.10 (-78.89%)	0.08 (-80.11%)	0.10 (-81.46%)

Table 5: F1 scores on context perturbed datasets. High F1 score and low percentage drops are desirable.

6.3 Context perturbation

Despite including mechanisms to incorporate contextual information, NER models show large performance drops under context perturbation attacks. Since an NER model is highly likely to have come across *"London"* as a LOCATION and *"Alice"* as a PERSON during training, it predicts them as such, ignoring the local context in which they appear. Despite large performance drops in general, Flair outperforms other models for all five datasets. This suggests that Flair captures local context better, likely due to the use of character embeddings.

Model	CoNLL	WIKI	GMB	FIRE	IEER
Flair	0.36 (-53.55%)	0.20 **(-69.33%)**	0.40 (-48.97%)	0.27 **(-64.31%)**	0.28 **(-63.50%)**
spaCy	0.27 (-52.94%)	0.19 (-67.20%)	0.38 (-49.02%)	0.26 (-61.03%)	0.27 (-61.09%)
CoreNLP-s	0.29 (-51.93%)	0.20 (-66.14%)	0.38 (-50.77%)	0.27 (-56.57%)	0.28 (-58.69%)
DeepPavlov	0.23 **(-55.00%)**	0.21 (-66.59%)	0.35 (-57.21%)	0.29 **(-51.13%)**	0.27 (-63.02%)

Table 6: F1 scores on context altered datasets. High percentage drops are desirable.

6.4 Context alteration

We note here that unlike the previous experiments, it is desirable to have higher percentage drops in performance for the context alteration attacks.[3] All the models show drops in performance. This hints at NER models having a tendency to learn the names themselves during training, rather than relying on the context in which the names appear. The magnitude of drops in performance is generally less than that observed for context perturbation, which suggests that NER models capture the local context of named entities better than their global context. Flair shows the largest performance drops, closely trailed by BERT.

7 Discussion

The adversarial evaluation of NLP models rely either on human-generated adversaries (Kaushik et al., 2019) or automated adversary generation with human-in-the-loop (Alzantot et al., 2018). However, it is possible to do away with human intervention for generating adversarial samples for the task of NER, as we demonstrate. Further, unlike existing work, our approach for adversarial evaluation does not require any re-training or fine-tuning of models for adversarial data creation.

The generalizability of NER models can also be evaluated with the proposed approaches. In particular, context perturbation can be used as an alternative to studying the effect of named entities that have not been seen during training (Augenstein et al., 2017) with the same label.

8 Conclusions

In this paper, we present an adversarial evaluation of four popular named-entity recognizers on five English datasets. The four model-agnostic adversarial attacks we present do not require white-box access to pre-trained NER models. Our experiments show that the popular NER models are over-reliant on the case information and under-utilise the contextual information. Since NER is a pre-requisite for a large number of NLP tasks, further work for improvement in these directions is warranted.

[3]Lower F1 scores are also desirable. However, low F1 scores can also be caused due to a model being poor generally and not specifically due to the inability to capture global context. Thus, we cannot draw concrete conclusions from the absolute F1 scores.

References

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

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

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Yuxuan Bai, Yu Wang, Bin Xia, Yun Li, and Ziye Zhu. 2020. Adversarial named entity recognition with pos label embedding. In *2020 International Joint Conference on Neural Networks*, pages 1–8. IEEE.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. 2009. Named entity recognition in wikipedia. In *Proceedings of the Workshop on The People's Web Meets NLP*, pages 10–18.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836.

Sravan 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*, pages 237–242.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In *Handbook of linguistic annotation*, pages 463–496.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yurii Kuratov, Denis Kuznetsov, et al. 2018. Deeppavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018*, pages 122–127.

Han-Cheol Cho et al. 2013. Named entity recognition with multiple segment representations. *Information Processing & Management*, 49(4):954–965.

Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 387–396.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 31–36.

Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

Matthew Honnibal. 2016. spacy.

Peixin Huang, Xiang Zhao, Minghao Hu, Yang Fang, Xinyi Li, and Weidong Xiao. 2022. Extract-select: A span selection framework for nested named entity recognition with generative adversarial training. In *Findings of the Association for Computational Linguistics*, pages 85–96.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

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

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

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

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics*, pages 55–60.

Stephen Mayhew, Gupta Nitish, and Dan Roth. 2020. Robust named entity recognition with truecasing pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8480–8487.

Stephen Mayhew, Tatiana Tsygankova, and Dan Roth. 2019. ner and pos when nothing is capitalized. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6257–6262.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. In *IEEE Military Communications Conference*, pages 49–54. IEEE.

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

Walter Simoncini and Gerasimos Spanakis. 2021. Seqattack: On adversarial attacks for named entity recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 308–318.

Tomasz Stanislawek, Anna Wróblewska, Alicja Wójcicka, Daniel Ziembicki, and Przemyslaw Biecek. 2019. Named entity recognition - is there a glass ceiling? In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 624–633.

Edward Loper Steven Bird and Ewan Klein. 2009. *Natural Language Processing with Python.*

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003*, pages 142–147.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2153–2162.

Jiuniu Wang, Wenjia Xu, Xingyu Fu, Guangluan Xu, and Yirong Wu. 2020. Astral: adversarial trained lstm-cnn for named entity recognition. *Knowledge-Based Systems*, 197:105842.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *6th International Conference on Learning Representations*.