Investigating the Vulnerability of Relation Extraction Models to Semantic Adversarial Attacks

Anonymous ACL submission

Abstract

In recent years, Large Language Models have set state-of-the-art performance on many NLP tasks. However, these models have been shown to be susceptible to permutations in data, and as such vulnerable to adversarial attacks.
In this work, we test the extent of this vulnerability with regards to models fine-tuned for the task of Relation Extraction by generating semantically-close adversarial samples using semantic information on relations, retrieved from an external knowledge base.
The results show that fine-tuned models for Relation Extraction are overall affected negatively by adversarials. Our results demonstrate that existing state-of-the-art Relation Extraction models are vulnerable to such adversarial attacks, with performance reductions of up to 33% in *F1* score, and with even the most robust model showing a decrease in *F1* score by 18%.
We also observe that certain patterns arise when the different models face specific permutations, regardless of the architecture implemented.

1 Introduction

Recent large language models (LLM) have achieved state-of-the-art performance on Relation Extraction (RE). They exploit contextual information from sentences to label relations between entity mentions (Baldini Soares et al., 2019; Wu and He, 2019). Despite achieving generally good results, these models show a lack of robustness when put under pressure (Tenney et al., 2020), such as under domain shift (Blitzer, 2008), and in adversarial settings (Papernot et al., 2016; Jia and Liang, 2017; Ebrahimi et al., 2018; Belinkov and Bisk, 2018).

In this work, we investigate the performance of these models in such an adversarial setting. In particular, we consider attacks that introduce changes which modify the semantics of a sentence while leaving the actual relation intact. Given a text example expressing a relation r between entities e_{subj} and e_{obj}, we replace e_{subj}, e_{obj}, or both entities by other entities. Consider the following sentence expressing a relation between *Leonardo Da Vinci* (e_{subj}) and the *Mona Lisa* (e_{obj}):

"*Leonardo Da Vinci painted the Mona Lisa.*"

Switching the entity *Leonardo Da Vinic* with other entities, such as *Michelangelo*, *Barack Obama* or *Stratolaunch* will arguably change the truth value (semantics) of the sentence, but a RE system should still recognize this relation to hold given that beyond the entities the further context is unchanged.

Our experiments with different types of substitution operations show that the models are significantly misled by such adversarials, reducing performance by between 18% and 33% in *F1* depending on the substitution operation.

2 Related Work

The construction of datasets capable of fooling neural NLP models has seen a surge in popularity in the last years, as proven by the *Build it, Break it* shared task (Bender et al., 2017), which encourages researchers to build linguistically-motivated examples to "break" NLP models in order to shed light over their weaknesses.[1] One way to build such datasets is through the generation of *adversarials*, i. e., examples altered through the addition of noise. The main idea is that a robust model should not be distracted by these adversarials (Jia and Liang, 2017).

The core idea behind the use of adversarials to test ML models has its roots in Computer Vision, where the incorporation of small perturbations into input images has been proposed as a way to create difficult-to-solve datasets (Szegedy et al., 2014; Goodfellow et al., 2015). Similar frameworks have been implemented for NLP tasks as well by em-

[1] https://bibinlp.umiacs.umd.edu/sharedtask.html

ploying different kinds of permutations.

Most of the approaches for adversarials have considered tasks other than RE, e. g., Li et al. (2016) use feature erasure to explain neural model decisions over several tasks, such as POS tagging and word frequency prediction. Similarly, Ribeiro et al. (2018) use adversarials to investigate bugs in machine comprehension, visual QA, and sentiment analysis. Further, Hosseini et al. (2017) have shown the effect of symbol addition and typo insertion on the task of toxicity detection, while Belinkov and Bisk (2018) applied such permutations to break machine translation.

While these models prove the effects of adversarials on NLP models, they lack in complexity and depth, as they exploit simple surface patterns (Wallace et al., 2019). A more complex framework in this regard is proposed by Li et al. (2021), who make use of enity-altering permutations to investigate the robustness of *BERT*-based models for RE. However, they only investigate the substitution of entities by entities of the same type and masks. Moreover, they only evaluate their adversarials on *BERT*-based models.

Beyond existing works and in particular the work of Li et al. (2021), we consider the impact of adversarials in RE by examining the effect of different type of entity substitutions, while also taking into account the performance of various state-of-the-art RE models.

3 Methodology

In our work, we aim at generating adversarials in which the semantics of the sentence is changed, while the type of the expressed relation stays the same. We argue that, while information about entities involved in the relation is crucial to determine the *truth value* of a sentence, this information should not play a critical role in predicting the relation label based on textual information.

For instance, given the example about the *Mona Lisa* in Sec. 1, while adversarial examples may describe false situations, they all express the same relation between `subj` and `obj`. Based on this assumption, we implement several substitution strategies to create a new adversarial dataset.

Given a corpus D as a set of quadruples of the form $(d, r, e_{subj}, e_{obj})$ where d is a sentence in which the relation r between the subject entity e_{subj} and the object entity e_{obj} is expressed, let E denote a set of all entities appearing in the dataset and let $E_{test} \subseteq E$ denote the set of entities that occur in the test subset of the corpus D_{test}, i. e., $E_{test} := \{e \in E \mid (d, r, e_{subj}, e_{obj}) \in D_{test} \wedge e \in \{e_{subj}, e_{obj}\}\}$. E_{train} and E_{val} are constructed analogously. Finally, let t_e denote the set of types of an entity e.[2]

Given a quadrupel $(d, r, e_{subj}, e_{obj}) \in D_{test}$, we replace e_{subj} or e_{obj} (or both) to obtain a new quadrupel that we then add to the set of adversarial examples D_{adv}. From which sets we randomly select an entity is defined by our substitution procedures, each allowing us to investigate a specific semantic phenomenon. The strategies are the following:

- **same-role substitution**: We obtain an adversarial example by replacing e_{subj} with e'_{subj} where e'_{subj} is randomly selected from the set $\{x \mid \exists(d', r, e'_{subj}, e'_{obj}) \in D_{train} : d \neq d'\}$. The entity e_{obj} can be replaced with e'_{obj}, analogously. Thus, an entity is replaced with another entity that occurs in the same role (as subject or object) with the same relation in another sentence in the training set.
- **same-type substitution**: We obtain an adversarial example by replacing e_{subj} with e'_{subj} where e'_{subj} is randomly selected from the set $\{x \in E_{train} \mid t_x \cap t_{e_{subj}} \neq \emptyset \wedge \neg\exists(d', r, y, e'_{obj}) : d \neq d' \wedge x = y\}$. The entity e_{obj} can be replaced with e'_{obj}, analogously. Thus, an entity is replaced with another entity occuring in the training set such that the original entity and the new entity have a common type and such that this entity never occured with that relation in that position in the training set.
- **different-type substitution**: We obtain an adversarial example by replacing e_{subj} with e'_{subj} where e'_{subj} is randomly selected from the set $\{x \in E_{train} \mid t_{e_{subj}} \cap t_x = \emptyset \wedge \neg\exists(d', r, x, y) : d \neq d'\}$. The entity e_{obj} can be replaced with e'_{obj}, analogously. Thus, an entity is replaced with another entity occurring in the training set such that the new entity does not have a type in common with e_{subj} and that does not occur in the same role in any document where the relation r is expressed.
- **masking**: We obtain an adversarial example by replacing e_{subj} with the `[MASK]` token. Analogously, we can replace e_{obj}.

We apply each of our four strategies to either replace the subject or the object or both to each

[2] We collect an entity's types by querying the Wikidata *SPARQL* endpoint: `https://query.wikidata.org/sparql`

element of the training set, thus obtaining 12 adversarial examples. The 12 adversarials generated for the *Mona Lisa* example introduced in Sec. 1 are shown in Fig. 1.

Original sentence:

(Leonardo da Vinci)$_{\text{subj}}$ painted the (Mona Lisa)$_{\text{obj}}$

Adversarials (subj mod.)

same-role:	Michelangelo$_{\text{subj}}$	painted the (Mona Lisa)$_{\text{obj}}$
same-type:	Barack Obama$_{\text{subj}}$	painted the (Mona Lisa)$_{\text{obj}}$
diff.-type:	Stratolaunch$_{\text{subj}}$	painted the (Mona Lisa)$_{\text{obj}}$
masking:	[MASK]$_{\text{subj}}$	painted the (Mona Lisa)$_{\text{obj}}$

Adversarials (obj mod.)

same-role:	(Leonardo da Vinci)$_{\text{subj}}$ painted the	Scream$_{\text{obj}}$
same-type:	(Leonardo da Vinci)$_{\text{subj}}$ painted the	Baloon Girl$_{\text{obj}}$
diff.-type:	(Leonardo da Vinci)$_{\text{subj}}$ painted the	Berlin Wall$_{\text{obj}}$
masking:	(Leonardo da Vinci)$_{\text{subj}}$ painted the	[MASK]$_{\text{obj}}$

Adversarials (subj and obj mod.)

same-role:	Michelangelo$_{\text{subj}}$	painted the	Scream$_{\text{obj}}$
same-type:	Barack Obama$_{\text{subj}}$	painted the	Baloon Girl$_{\text{obj}}$
diff.-type:	Stratolaunch$_{\text{subj}}$	painted the	Berlin Wall$_{\text{obj}}$
masking:	[MASK]$_{\text{subj}}$	painted the	[MASK]$_{\text{obj}}$

Figure 1: Examples of generated adversarials

4 Experiment

4.1 Data

The starting point of our adversarial dataset is FewRel[3] (Han et al., 2018; Gao et al., 2019), a large few-shot RE dataset created through a combination of distant supervision and human annotation. While the original objective of FewRel is to train models on few-shot RE, the original data can be used to generate a new dataset that can be used in a standard RE framework.

In order to create our own dataset, we combine the `train` and `val` datasets of FewRel, containing respectively 64 and 16 relations, each with 700 examples, for a total of 56,000 sentences. For each entity mention, the dataset also contains their corresponding Wikidata entity ID. For the purpose of this work, the combined dataset is randomly split into train/test/dev splits with percentages 70/15/15. Starting from the test split, we generate adversarial examples as explained in Section 3.[4]

4.2 Training parameters

The investigated models are fine-tuned on the original dataset using either static or contextual embeddings. For static embeddings, we use the 100-dimensional word vectors proposed in (Turian et al., 2010), while for contextual embeddings we employ the 768-dimensional `bert-base-uncased` model for English (Devlin et al., 2019)[5].

The first text-based model is an biLSTM + Attention model (*ATT-biLSTM*) based on the work by Zhou et al. (2016). The model is trained for 50 epochs.

We further test a convolutional neural network (*CNN*), inspired by (Zeng et al., 2014), which feeds the concatenation of entities', sentences' and positional vectors to a convolutional layer. This model is trained for 20 epochs.[6]

The third model is BERT with entity markers (*BERT$_{em}$*) (Baldini Soares et al., 2019), which uses entity tags inserted before and after each entity to represent relations in vector space. This model is fine-tuned for 5 epochs.[7]

The fourth model is *R-BERT* (Wu and He, 2019), which concatenates the embeddings for the `[CLS]` token and the average pooling of entity mentions' token, and then feeds the result to a fully connected layer. This model is fine-tuned for 5 epochs.[8]

The final model is represented by RIFRE, proposed by Zhao et al. (2021). This model makes use of an heterogeneous graph representation, composed of word nodes and relation nodes, which is used to update word representations. This model is fine-tuned for 5 epochs.[9]

5 Results

Once fine-tuned, the models are tested on standard and adversarial examples, and evaluated using the F1 score, as shown in Table 1. In the experiment, the best results are achieved by the *RIFRE* model, on both the standard test set and the adversarial one, with a standard $F1$ of 0.9 and an adversarial $F1$ of 0.72. Even though the *BERT_em* model performs worse on the normal test set, its scores on the adversarial test set are comparable to the scores achieved by the *RIFRE* model. As such,

[3] https://github.com/thunlp/FewRel

[4] We provide the adversarial dataset for download at https://www.dropbox.com/scl/fo/yiutamdzbbe7g8oe7hnww/h?dl=0&rlkey=fi6qv9wpodzwr5u6pdnvwuq51

[5] Unless stated otherwise, all the hyperparameters are set according to the original papers

[6] We used the unofficial implementation found at https://github.com/onehaitao/CNN-relation-extraction

[7] We used the unofficial implementation found at https://github.com/plkmo/BERT-Relation-Extraction

[8] We used the unofficial implementation found at https://github.com/monologg/R-BERT

[9] We used the official implementation found at https://github.com/zhao9797/RIFRE

Table 1: Fine-grained F1 scores on *FewRel (custom)*. **std** contains standard evaluation, **adv** adversarial evaluation, **diff** the difference between the two. The following columns contain $F1$ score for all the strategies.

Model	**std.**	**adv.**	**diff**	**same-role sub.**			**same-type sub.**			**diff.-type sub.**			**masks sub.**		
				subj	obj	subj + obj	subj	obj	subj + obj	subj	obj	subj + obj	subj	obj	subj + obj
Att-BiLSTM	.66	.44	-33%	.68	.62	.41	.65	.58	.57	.55	.28	.20	.51	.16	.06
CNN	.71	.56	-21%	.72	.69	.50	.71	.67	.66	.58	.38	.26	.64	.49	.41
R-BERT	.89	.7	-21%	.88	.86	.57	.87	.84	.83	.74	.43	.30	.83	.66	**.56**
RIFRE	**.90**	**.7**	-20%	**.89**	**.87**	.62	**.88**	**.85**	**.83**	**.77**	**.54**	**.39**	**.85**	**.73**	.51
BERT_em	.84	.69	**-18%**	.85	.83	**.64**	.84	.81	.79	.72	**.54**	.37	.75	.63	.53
avg	.75	.54	-28%	.75	.71	.45	.73	.67	.65	.59	.36	.22	.61	.43	.28

the *BERT_em* is the most robust model under our adversarial attacks, as its scores decrease only by 18%, whereas the scores of the other models are more strongly impacted by the adversarials. The worst performing model under adversarials is the *ATT-biLSTM* model, for which the scores decrease by 33%.

The investigation of results per substitution strategy unveils patterns that are found across all models, proving that different models react in similar way to certain permutations.

Across all substitution strategies, the substitution of the *subject* entity mention does not have as strong of an impact as the substitution of the *object* entity mention on the *F1* scores.

This is particularly evident for the **diff.-type substitution** and **masks** substitutions, with the former showing an *F1* of 0.59 when the subject is substituted, and 0.36 when the object is substituted. Similarly, mask substitution has an *F1* of 0.61 when subject is substituted, and 0.43 when the substitution affects the object.

Secondly, while the first three substitution strategies show a decreasing level of performance, with **diff.-type substitution** showing worse results than **same-type substitution**, and **same-type substitution** showing worse results than **same-role substitution**, it is interesting to notice that **masking** does not show worse results than **diff.-type substitution**. Actually, in certain cases, masking an entity shows improvements over randomly substituting it, showing that models rely too much on entity mentions to predict relations' label, rather than the actual realization of relations.

6 Conclusion and Future Work

The experiments described in this work show the vulnerability of current RE approaches to semantically similar adversarials. Overall, the performance of the examined models shows a substantial degradation of 18% and 33% in *F1*. In particular, we found that the substitution of both entities with entities of different type has the worst effect on models, with an average *F1* score of 0.22. Our evaluation also shows patterns that are found across all models, proving that different models react in similar way to certain permutations. For instance, the substitution of the object entity usually has a stronger impact than the substitution of the subject entity.

In future works, we aim at investigating methodologies for improving models' resistance to such attacks. Furthermore, the current investigation could be made more precise through the inclusion of similarity values between different relations' realization as a weighted component in the final evaluation of the models.

7 Limitations

We realize that one of the limitations of the proposed approach is that the adversarial substitution might result in false negatives, in those cases in which the relations are expressed through similar structures. Given, for instance, the sentence

"*Leonardo da Vinci was born in Anchiano*"

and a possible adversarial example

"*Leonardo da Vinci was born in 1452*"

The example does not retain its original relation label, hence affecting the final evaluation. In order to investigate the amount of possible false negatives, we compute the average token overlap between relations' examples by implementing the Szymkiewicz–Simpson coefficient (M.K and Kavitha, 2016). Since sentences referring to different relations have, in general, an average of one fifth of their tokens in commonit can be assumed that the actual overlap between sentences expressing different relation is low enough for false negatives not to be considered an issue, and that these cases are the exceptions rather than the rule.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Emily M. Bender, Hal Daumé III, Allyson Ettinger, Harita Kannan, Sudha Rao, and Ephraim Rotschild. 2017. Build, break it: The language edition.

John Blitzer. 2008. *Domain Adaptation of Natural Language Processing Systems*. Ph.D. thesis, USA. AAI3309400.

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.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

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.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure.

Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. On robustness and bias analysis of bert-based relation extraction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, pages 43–59, Singapore. Springer Singapore.

Vijaymeena M.K and K. Kavitha. 2016. A survey on similarity measures in text mining.

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

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118, Online. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, page 384–394, USA. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2361–2364, New York, NY, USA. Association for Computing Machinery.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Kang Zhao, Hua Xu, Yue Cheng, Xiaoteng Li, and Kai Gao. 2021. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction. *Knowledge-Based Systems*, 219:106888.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.